UNIVERSITY OF CALIFORNIA, SAN DIEGO

The Genetics and Epigenetics of Induced Pluripotent Stem Cells

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in

Bioengineering

by

Athurva Jayavant Gore

Committee in charge:

Professor Kun Zhang, Chair
Professor Adam Engler
Professor Lawrence S. B. Goldstein
Professor Xiaohua Huang
Professor Alysson Muotri

2013

The Dissertation of Athurva Jayavant Gore is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

Chair

University of California, San Diego

2013

DEDICATION

For my wife, my parents, and my brother. Thank you for all your support.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to thank everyone who supported me in ways both big and small. Without the support of the following people, this would not be possible.

First, I would like to thank my advisor, Kun Zhang. I am very glad that he was able to meet with me on short notice when I visited UCSD as a prospective graduate student. He offered me a rotation in his laboratory based just on that short meeting, and his support has allowed me to succeed. He gave me the freedom to explore a multitude of different projects, all of which were exciting. His support has helped me reach where I am today.

I would also like to thank the rest of my committee members: Dr. Adam Engler, Dr. Lawrence Goldstein, Dr. Xiaohua Huang, and Dr. Alysson Muotri. They were always available for advice when asked, and gave me valuable feedback throughout my time as a graduate student researcher.

I would also like to thank everyone in the Zhang lab for supporting me throughout my time as a graduate student. I would like to thank Alan (Ho-Lim) Fung for keeping the sequencer up and running and Dinh Diep, Alice (Zhe) Li, Rui Liu, and Noi (Nongluk) Plongthongkum for their help with experiments. I would specifically also like to thank Jeff Gole for all of the helpful advice over lunch.

Finally, I would like to thank my family. My wife, parents, and brother have been there for me throughout my graduate school experience. Without their support, I would have struggled greatly.

Chapter 2, in part, is a reprint of the material as it appears in: Dinh Diep*, Nongluk Plongthongkum*, Athurva Gore*, Ho-Lim Fung, Robert Shoemaker, Kun Zhang. “Library-free Methylation Sequencing with Bisulfite Padlock Probes.” *Nature Methods*. 2012 February 5; 9(3): 270-272. doi: 10.1038/nmeth.1871. Used with

permission. The dissertation author was one of the primary investigators and authors of this paper.

Chapter 3, in part, is a reprint of the material as it appears in: Athurva Gore*, Zhe Li*, Ho-Lim Fung, Jessica E. Young, Suneet Agarwal, Jessica Antosiewicz-Bourget, Isabel Canto, Alessandra Giorgetti, Mason A. Israel, Evangelos Kiskinis, Je-Hyuk Lee, Yuin-Han Loh, Philip D. Manos, Nuria Montserrat, Athanasia D. Panopoulos, Sergio Ruiz, Melissa L. Wilbert, Junying Yu, Ewen F. Kirkness, Juan Carlos Izpisua Belmonte, Derrick J. Rossi, James A. Thomson, Kevin Eggan, George Q. Daley, Lawrence S. B. Goldstein, Kun Zhang. "Somatic coding mutations in human induced pluripotent stem cells." *Nature*. 2011 March 3; 471: 63-67. doi:10.1038/nature09805. Used with permission. The dissertation author was one of the primary investigators and authors of this paper.

Chapter 4, in part, is a reprint of the material as it appears in: Sergio Ruiz*, Athurva Gore*, Zhe Li, Athanasia D. Panopoulos, Nuria Montserrat, Ho-Lim Fung, Alessandra Giorgetti, Josipa Bilic, Erika M. Batchelder, Holm Zaehres, Hans R. Scholer, Kun Zhang, and Juan Carlos Izpisua Belmonte. "Analysis of protein-coding mutations in hiPSCs and their possible role during somatic cell reprogramming." *Nature Communications*. 2013 January 22; 4: 1382. doi:10.1038/ncomms2381. Used with permission. The dissertation author was one of the primary investigators and authors of this paper.

Chapter 6, in part, is a reprint of the material as it appears in: Sergio Ruiz*, Dinh Diep*, Athurva Gore, Athanasia D. Panopoulos, Nuria Monsterrat, Nongluk Plongthongkum, Sachin Kumar, Ho-Lim Fung, Alessandra Giorgetti, Josipa Bilic, Erika M. Batchelder, Holm Zaehres, Natalia G. Kan, Hans Robert Scholer, Mark Mercola, Kun Zhang, Juan Carlos Izpisua Belmonte. "Identification of a specific

reprogramming-associated epigenetic signature in human induced pluripotent stem cells." *PNAS*. 2012 October 2; 109 (40): 16196-16201. doi:10.1073/pnas.1202352109. Used with permission. The dissertation author was one of the primary investigators and authors of this paper.

VITA

2007 Bachelor of Science, Biomedical Engineering, Purdue University

2013 Doctor of Philosophy, Bioengineering, University of California, San Diego

PUBLICATIONS

Deng, J. et al. Targeted bisulfite sequencing reveals changes in DNA methylation associated with nuclear reprogramming. Nature biotechnology 27, 353-360 (2009).

Gore, A. et al. Somatic coding mutations in human induced pluripotent stem cells. Nature 471, 63-67 (2011).

Howden, S.E. et al. Genetic correction and analysis of induced pluripotent stem cells from a patient with gyrate atrophy. Proceedings of the National Academy of Sciences 108, 6537-6542 (2011).

Noggle, S. et al. Human oocytes reprogram somatic cells to a pluripotent state. Nature 478, 70-75 (2011).

Wang, X. et al. Whole exome sequencing identifies ALMS1, IQCB1, CNGA3, and MYO7A mutations in patients with leber congenital amaurosis. Human mutation 32, 1450-1459 (2011).

Diep, D. et al. Library-free methylation sequencing with bisulfite padlock probes. Nature methods 9, 270-272 (2012).

Ball, M.P. et al. A public resource facilitating clinical use of genomes. Proceedings of the National Academy of Sciences 109, 11920-11927 (2012).

Ruiz, S. et al. Identification of a specific reprogramming-associated epigenetic signature in human induced pluripotent stem cells. Proceedings of the National Academy of Sciences 109, 16196-16201 (2012).

Michaelson, J.J. et al. Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. Cell 151, 1431-1442 (2012).

Ruiz, S. et al. Analysis of protein-coding mutations in hiPSCs and their possible role during somatic cell reprogramming. Nature communications 4, 1382 (2013).

Gole, J. et al. Massively parallel polymerase cloning and genome sequencing of single cells using nanoliter microwells. Nature biotechnology (2013).

Woodruff, G. et al. The Presenilin-1 ΔE9 Mutation Results in Reduced γ-Secretase Activity, but Not Total Loss of PS1 Function, in Isogenic Human Stem Cells. Cell reports (2013).

FIELDS OF STUDY

Major Field: Bioengineering

Studies in High Throughput Genomics and Epigenomics
Professor Kun Zhang

Studies in Stem Cells and Induced Pluripotency
Professor Kun Zhang

ABSTRACT OF THE DISSERTATION

The Genetics and Epigenetics of Induced Pluripotent Stem Cells

by

Athurva Jayavant Gore

Doctor of Philosophy in Bioengineering

University of California, San Diego, 2013

Professor Kun Zhang, Chair

The ability to induce pluripotency in human adult somatic cells by defined transcription factor expression is a revolutionary prospect in regenerative medicine. This discovery has the potential to both open new research avenues for diseases in tissue types that are difficult to obtain and to revolutionize medicine through the use of patient-derived replacement tissue. However, questions remain about the safety and efficacy of these induced pluripotent stem cells (iPSCs). Because iPSC generation protocols tend to be low efficiency, require derivation from adult tissue, often utilize viral transfection, force the expression of known oncogenes, and involve a large number of rapid cell divisions during reprogramming, it was thought that the iPSC genome itself might contain some genetic mutation. Additionally, the progenitor

cell type used for iPSC derivation seemed to cause some differentiation pathways to be more highly favored, indicating that iPSCs might possess some sort of “epigenetic memory” of their progenitor state.

Thanks to modern advances in high throughput sequencing, we were able to assess the genomic and epigenomic state of induced pluripotent stem cells, and thus determine if iPSCs could be used in either a clinical or a research context. We demonstrate that induced pluripotent stem cells contain a large number of point mutations across their genome regardless of donor age, time in culture, progenitor cell type, or reprogramming method. While a majority of these mutations arise due to rare progenitor mutations becoming fixed through clonal selection during reprogramming, approximately 43% arise either during the reprogramming step or during iPSC expansion. We additionally show that, in addition to epigenetic memory of the progenitor cell state and aberrant DNA methylation, nearly all iPSC lines carry a unique reprogramming-specific epigenetic signature that remains even after further differentiation and impacts gene expression in iPSC-derived cells. Taken together, these results demonstrate that iPSCs must still overcome major hurdles prior to their widespread clinical use. Rigorous work towards establishing clinical safety standards for genetic and epigenetic integrity in pluripotent-derived therapies will be essential before the promise of induced pluripotency can be fully realized.

Chapter 1: Introduction

1.1 Next Generation Sequencing

The field of DNA sequencing has undergone a drastic shift over the last eight years. The advent of next generation sequencing and its rapid adoption in a research setting have led to the development of new protocols and the characterization of genetic variation on a scale that was prohibitively expensive a decade ago[1]. Modern sequencing instruments can generate up to 600 billion bases of sequencing data in a single run, allowing an entire human genome to be sequenced at high depth in only two weeks. This is a vast improvement over the original Sanger sequenced human genome, which took over a decade to sequence and analyze[2]. The cost of sequencing has also greatly decreased at a rate orders of magnitude greater than predicted by Moore's law; while the original human genome cost nearly $3 billion to sequence, an entire human genome can now be sequenced for between $5,000 and $10,000[3]. Additionally, the commercialization of high-throughput sequencers by companies such as Illumina, 454, and Life Technologies and the prevalence of "sequencing-as-a-service" core facilities have made sequencing accessible to nearly all engineers and scientists interested in genome analysis[1]. The Illumina GAIIx and HiSeq are currently the most popular sequencing instruments due to their large amount of generated data and low cost per run[1].

In order to perform a next generation sequencing experiment on the Illumina platform, a "library construction" protocol must be performed on the input DNA. In this process, input DNA is fragmented into smaller pieces, typically between 200-800 base pairs, and specific oligonucleotide adapter sequences are added to the ends of

each molecule[4]. These sequences can also include sample-specific barcodes, allowing multiple samples to be processed during a single sequencing run. For classical library construction, template DNA is generally sheared using a sonication device such as a Covaris Adaptive Focused Acoustics machine, which produces high-energy waves that fragment DNA to a tight size range[4]. After this process, DNA must be end repaired and A-tailed; adapters can then be ligated using a TA-ligation process. PCR is then used to amplify each individual molecule and generate the finished sequencing library. The library must then be quantified for accurate dilution and loading into the sequencer[5]. Before each experimental step, a purification step is carried out in order to ensure optimal reaction conditions; however, this results in compounding losses of template DNA. Thus, hundreds of thousands of input cells (with micrograms of genomic DNA) are necessary to properly construct a sequencing library using this method[4].

To overcome the need for a large number of input cells and many experimental and purification steps, an alternative library construction technique based on random transposition was developed. Known as Nextera, this method relies on a modified transposase enzyme loaded with a transposon containing Illumina sequencing adapters[6]. Unlike classical transposase reactions, in a Nextera reaction, a gap is present in the inserted transposon; because of this, the transposition reaction will both randomly fragment and insert sequencing adaptors into template DNA in one step. PCR can be directly performed on the transposed molecules to create a sequencing library without any intermediate purification. Researchers have used Nextera to generate successful sequencing libraries from as little as 10 picograms of genomic DNA (~2 human cells)[6]. Nextera therefore provides an alternative to classical library construction when smaller amounts of template DNA

are available, such as when analyzing DNA from rare organisms or from limited whole genome amplification of a few cells.

During a sequencing run, the constructed library is denatured into single-stranded DNA, diluted, and loaded into a flowcell containing clusters of affixed small DNA molecules complimentary to the adapter sequences. The loading concentration is calibrated such that each individual input molecule will anneal relatively far apart from all others in the flowcell[1]. A modified form of PCR is then performed to amplify each individual molecule into a cluster of copies using the nearby affixed DNA as primers; Illumina's sequencing-by-synthesis protocol can then be used to obtain the DNA sequence of each molecule. Due to error and drift effects during synthesis, Illumina sequencers can only sequence ~200 base pairs from each end of each input fragment; however, this amount is sufficient for resequencing experiments[5].

1.2 Targeted Sequencing

Despite the rapidly decreasing cost of sequencing, it is not yet feasible to sequence and analyze large numbers of whole genomes. Performing a large experiment on hundreds of samples simultaneously, soon to be a common requirement for in-depth sequencing-based disease studies, would still cost hundreds of thousands of dollars; analysis time and data storage would also be very difficult for a small research laboratory[7]. In addition, many disease or mutation-related studies only require investigation of small portions of the genome, such as genes known to be causative in disease or protein-coding regions; performing whole-genome sequencing in these experiments is unnecessary.

Because of these issues, multiple targeted sequencing methods have been developed to allow analysis of selected regions of the genome. One of the simplest

and most commonly used methods of target enrichment utilizes polymerase chain reaction, or PCR[8]. By utilizing two PCR primers around a region of interest with sequencing adapters already included as 5' primer overhangs, single sections of the genome can easily be converted to a sequencing library format. However, due to cross-primer interactions and difficulties in ensuring uniform amplification between multiple primer sets, multiplex PCR is extremely difficult to perform on a large scale; to target more than a few genomic regions, PCR reactions must either be run in many separate uniplex reactions or utilize another method of reaction segregation such as emulsion-based PCR, in which every droplet in a reaction contains a separate set of unique primers[9]. For large target regions, even these approaches are insufficient.

Another approach to targeted sequencing relying on a similar principle to PCR is molecular inversion probes or padlock probes. Instead of utilizing two separate primers to amplify regions of interest exponentially using both DNA strands, padlock probes involve utilizing several sets of two primers joined together by a common linker sequence containing sequencing adaptors[10]. During a reaction, the two primers anneal upstream and downstream of a region of interest on the same DNA strand; the gap between them is filled using a polymerase, and the probe molecules are then circularized. The linked upstream and downstream primers prevent cross-primer interactions, and use of circularization additionally improves targeting specificity. However, probes have been reported to have widely different capturing efficiencies, meaning that certain regions might be difficult to capture[11]. Additionally, padlock probes have previously been relatively costly to synthesize due to the relatively high cost of multiple column-based synthesis reactions and relative unavailability of programmable DNA arrays.

Another method utilized to perform targeted sequencing is hybridization-based capture. In this method, many DNA fragments complimentary to specific regions of interest are synthesized and biotinylated. These fragments are annealed to a whole-genome sequencing library, and the biotinylated hybridized molecules are captured with streptavidin beads[12]. While hybridization capture is a highly scalable and efficient capturing procedure with more uniform capturing efficiency between different targets than padlock probes, cost-effective synthesis of capturing fragments can be an issue just as with padlock probes. Additionally, hybridization capture tends to be biased towards smaller DNA fragments in the sequencing library, and can also be inefficient at the extremes of high A+T and G+C content[13]. 5-15% of a desired target region can additionally be lost due to presence of repetitive regions, and many off-target regions will be captured due to incorrect hybridization during capture[12].

1.3 Bisulfite Sequencing

Eukaryotic cells regulate their gene expression profiles by many mechanisms, one of which is known as DNA methylation. DNA methylation generally refers to the presence of a methyl group (CH3) covalently bonded to the fifth carbon atom of cytosine[14]; this particular DNA modification has been implicated in regulation of gene expression, X-chromosome inactivation, genomic imprinting, and silencing of genes during cell differentiation. DNA methylation is often perturbed during human diseases such as imprinting disorders and cancer[15]. Because of this, understanding methylation patterns throughout different cell types is extremely important.

DNA methylation can be difficult to detect, as many enzymes and compounds do not distinguish between cytosine and methylated cytosine. Several methods have been developed to characterize DNA methylation, including enrichment of DNA

fragments with methyl binding proteins and extraction of DNA fragments after digestion with methylation-specific restriction enzymes[16, 17]. However, the best method to measure DNA methylation is bisulfite sequencing. Bisulfite sequencing involves treating genomic DNA with sodium bisulfite, a compound that converts cytosine to uracil but leaves methylated cytosine intact[18]. While this simple treatment with sodium bisulfite allows detection of DNA methylation at single-base resolution, it also presents several challenges. Bisulfite treatment must be performed on unamplified genomic DNA, as DNA polymerase recognizes methyl-cytosine as cytosine and therefore does not conserve DNA methylation[17]. Sodium bisulfite also tends to cause DNA damage, including double stranded breaks in DNA. Computational analysis of bisulfite data can additionally be very challenging; the removal of cytosine residues from the genome creates a very degenerate and repetitive reference sequence[14]. Because of these issues, large amounts of input DNA and large amounts of sequencing are required for a bisulfite sequencing experiment, often making the cost of performing a large-scale experiment prohibitive.

DNA methylation typically occurs in a CpG dinucleotide context, with both cytosines in the dinucleotide remaining methylated; very few non-CpG cytosines are methylated in differentiated tissues[19]. CpGs additionally tend to occur in dense patches throughout the genome known as CpG islands. Thus, because for the most part only CpG dinucleotides need to be analyzed for methylation, use of a targeted sequencing protocol can greatly improve the cost-effectiveness of bisulfite sequencing, allowing simultaneous processing of many different cellular samples[11].

1.4 Induced Pluripotent Stem Cells

Pluripotent cells, or those than can generate specific adult cell types from each of the three embryonic germ layers, have inspired the field of personalized regenerative medicine. Pluripotent cells allow the possibility of modeling and discovering mechanisms behind complex human diseases in a dish, and have the potential to enable cell therapy for previously untreatable conditions[20]. However, generation of pluripotent cells previously required the use of human embryos for derivation, leading to political controversy. Additionally, any cell therapies derived from these embryonic stem cells (or ESCs) would still face the same issues as adult cell or organ-based therapies in terms of immune rejection[21]. Researchers thought to utilize somatic cell nuclear transfer (SCNT) to address these issues and generate patient-specific pluripotent stem cell lines from adult cells using oocytes; however, while these experiments have been highly successful in other animals, generation of a diploid human SCNT line has remained elusive[22].

However, seven years ago, a new possibility was brought to the field of personalized regenerative medicine: induced pluripotent stem cells, or iPSCs. These pluripotent cells could be generated from mouse fibroblasts through the forced expression of only four genes (OCT4, SOX2, KLF4, and cMYC), and could then be differentiated into potentially any adult cell type[23]. This seminal finding was quickly translated to human fibroblasts and then to a variety of other widely available cell types. iPSCs have been differentiated into many different cell types, in some cases allowing the study of diseases in tissue types considered extremely difficult to obtain and unculturable[21]. It appeared that personalized regenerative medicine might truly be possible, as iPSC-derived cells in theory would not provoke the immune response

expected in ESC-based therapies, and would not require the use of human oocytes or embryos for generation.

However, while iPSCs appeared to demonstrate full pluripotency (especially in mice, where some iPSC lines were even used in tetraploid complementation experiments to generate full adult mice[24]), it became clear that many iPSC lines seemed to favor differentiation along certain lineages and differentiate into others less efficiently[25]. It appeared that the progenitor cell type utilized might influence which differentiation pathways were favored, indicating that iPSCs might possess some sort of "epigenetic memory" of their progenitor state[26]. In addition, because iPSC generation protocols were low efficiency, required derivation from adult tissue, often utilized viral transfection, forced the expression of known oncogenes, and involved a large number of rapid cell divisions during reprogramming, it was thought that the iPSC genome itself might contain some genetic mutation[27]. Thanks to modern advances in high throughput sequencing, it was possible to assess the genomic and epigenomic state of induced pluripotent stem cells, and thus determine if iPSCs could be used in either a clinical or a research context.

1.5 Scope of the Dissertation

The purpose of this dissertation was to characterize induced pluripotent stem cells at both a genetic and an epigenetic level. While iPSCs represent a large step forward in personalized regenerative medicine, in order to remain useful for both research and clinical applications, they must retain a stable genomic state and undergo complete epigenetic reprogramming to a pluripotent state. By characterizing these states, the unique challenges faced when using iPSCs can be better understood and compensated for.

In Chapter 2, we describe the refinement of a targeted sequencing method known as “padlock probes.” We describe the design and implementation of a padlock probe designer, an extremely fast and parallel sequencing library construction protocol, and a computational pipeline to analyze padlock probe data. We demonstrate that padlock probes provide an extremely scalable and robust way to perform targeted sequencing on many samples at low cost.

In Chapter 3, we describe the use of our padlock probe platform and other targeted sequencing platforms to identify somatic coding mutations in human induced pluripotent stem cells. We show that iPSCs contain an alarming number of mutations in protein-coding regions that might cause unpredictable behavior and negatively impact their usability in a clinical or research context.

In Chapter 4, we further characterize coding mutations in induced pluripotent stem cells. We demonstrate that point mutations are present in iPSCs regardless of the progenitor cell type used for derivation. We also show that point mutations in coding regions individually do not seem to favor the process of reprogramming in a loss-of-function context; mutations therefore appear to be more random in nature, and might show unpredictable behavior in iPSC-derived tissue.

In Chapter 5, we characterize the origin of somatic point mutations in human induced pluripotent stem cells. We perform a unique high-depth whole genome sequencing experiment in which we characterize the mutational load of iPSC lines at multiple passages, including at a 1000 cell pre-passaging state. We demonstrate that reprogramming-associated mutations arise in three separate categories, and that a majority of iPSC mutations pre-exist at low levels in the progenitor cells.

Lastly, in chapter 6, we characterize the epigenetics of induced pluripotent stem cells using our padlock probe platform. We show that iPSCs contain a wide

variety of aberrantly methylated regions throughout the genome, and that 9 specific regions appear to be aberrantly methylated in all iPSC lines generated by a multitude of research groups. We additionally demonstrate that these aberrant epigenetic patterns remain even after further differentiation, showing that iPSC-derived tissue exhibits aberrant gene expression patterns compared to ESC-derived tissue.

Chapter 2: Library-Free Bisulfite Sequencing with Padlock Probes

2.1 Abstract

We previously developed padlock probes (PPs) for the specific and parallel targeted resequencing of important portions of the genome[10] and digital quantification of DNA methylation[11]. In this chapter, we report the second-generation of padlock probe-based targeted sequencing with a design algorithm to generate more efficient padlock probes, a library-free protocol that dramatically reduces the time and cost of sample preparation and is compatible with automation, and an efficient bioinformatics pipeline that can accurately characterize both genomic variation and DNA methylation levels.

2.2 Introduction

DNA methylation is a widespread epigenetic mechanism by which vertebrate cells regulate gene expression[14]. In adult cells, it typically occurs in a CpG dinucleotide context, though non-CpG methylation has been observed in developing cells; in both cases, a methyl group is added to the fifth carbon of cytosine. DNA methylation is associated with suppression of gene expression. It is known to be aberrant in many diseases, including cancer and Rett syndrome[15]; thus, there is a need for characterization of this method of gene expression.

Many methods have been utilized to characterize DNA methylation across the genome[16, 17], including microarray hybridization, methylated DNA

immunoprecipitation, and methylation-specific restriction enzyme digestion. However, in order to quantify DNA methylation levels at single-nucleotide resolution, the best method is known as bisulfite sequencing[18]. In a bisulfite sequencing experiment, genomic DNA is treated with sodium bisulfite, causing all unmethylated cytosines to be converted to thymine; only methylated cytosines remain. Bisulfite sequencing has been previously utilized to accurately characterize methylation levels in multiple species[28, 29].

Although bisulfite sequencing theoretically allows characterization of methylation throughout the entire genome, the cost of such an experiment is extremely high. Because bisulfite treatment changes the majority of the genome into a 3-base system, an enormous amount of DNA sequencing is required. Many sequencing resources go to waste, as very few non-CpG dinucleotides are methylated. In terms of scalability to hundreds or thousands of samples, whole genome bisulfite sequencing remains prohibitively expensive[14].

Because of this, in order to reduce sequencing cost and improve experimental scalability, there is a need for a method to perform selection or enrichment of genomic targets prior to sequencing. One previously utilized method is PCR-based target amplification; while feasible for small numbers of targets, this method cannot easily be multiplexed to cover large portions of the genome[7]. Another method is reduced representation bisulfite sequencing (RRBS), in which a sample is digested with methylation-specific restriction enzymes and only certain sized products are sequenced; however, this method tends to bias sequencing towards CpG-dense areas and ignores many important genomic regions that have lower methylation levels overall[14].

We have addressed these limitations in targeted sequencing by using padlock probes to generate high-throughput sequencing libraries for an arbitrary set of sequencing targets. In this chapter, we describe a probe design algorithm capable of generating more efficient padlock probes, a library-free protocol that dramatically reduces the time and cost of sample preparation and is compatible with automation, and an efficient bioinformatics pipeline that can accurately characterize both genomic variation and DNA methylation levels. While we primarily describe the use of this pipeline to characterize DNA methylation in a set of arbitrary targets across the genome, we have also utilized this pipeline to characterize genomic variation across multiple cell lines. Sets of padlock probes have been used to characterize a variety of genomic regions from *Homo sapiens*[30, 31], *Mus musculus*[32], and *Drosophila melanogaster*[33] in both a genomic and epigenomic context.

2.3 Methods

Algorithms, probe sequences, and additional supplementary information are available at http://genome-tech.ucsd.edu/public/Gen2_BSPP/. A schematic of the padlock probes is illustrated in Supplementary Fig. 2.10.

2.3.1 Bisulfite padlock probe production (Agilent)

Libraries of oligonucleotides (~150 nt) were synthesized by ink-jet printing on programmable microarrays (Agilent Technologies) and released to form a combined library of 330,000 oligonucleotides. The oligonucleotides were amplified by PCR in 96 reactions (100 µl each) with 0.02 nM template oligonucleotide, 400 nM each of pAP1V61U primer and AP2V6 primer (Supplementary Table 2.3), and 50 µl of KAPA SYBR fast Universal 2x qPCR Master Mix (Kapa Biosystems) at 95 °C for 30 s, 15-16 cycles of 95 °C for 3 s; 55 °C for 30 s; and 60 °C for 20 s, and 60 °C for 2 min. The

amplicons were purified by ethanol precipitation and re-purified with Qiaquick PCR purification columns (Qiagen). Approximately 20 µg of the purified amplicons were digested with 50 units Lambda Exonuclease (5 U/ µl; New England Biolabs (NEB)) at 37 °C for 1 h in lambda exonuclease reaction buffer. The resulting single-stranded amplicons were purified with Qiaquick PCR purification columns. Approximately 5-8 µg of single stranded amplicons were subsequently digested with 5 units USER (1 U/µl, NEB) at 37 °C for 1 h. The digested DNA molecules were annealed to 5.88 µM RE-DpnII-V6 guide oligonucleotides (Supplementary Table 2.3) after initial denaturing at 94 °C for 2 min; the temperature was then slowly decreased to 37 °C and held for 3 min. The mixture was digested with 50 units DpnII (10U/µl, NEB) in NEBuffer DpnII at 37 °C for 2 h. The mixture was then further digested with 5 units USER at 37 °C for 2 h; enzymes were then inactivated through incubation at 75 °C for 20 min. The USER/DpnII digested DNA was then purified with Qiaquick PCR purification columns. The single-stranded probes (of 102 bases in length) were purified with 6% denaturing PAGE (6% TB-urea 2D gel; Invitrogen).

2.3.2 Bisulfite padlock probe production (LC Sciences)

Oligonucleotides (100 nt) were synthesized using a programmable microfluidic microarray platform (LC Sciences) and released to form a mix of 3,918 oligonucleotides. The oligonucleotides were amplified by two-step PCR in a 200 µl reaction with 1nM template oligonucleotides, 400 nM each of eMIP_CA1_F primer and eMIP_CA1_R primer (Supplementary Table 2.3), and 100 µl of KAPA SYBR fast Universal qPCR Master Mix at 95 °C for 30 s, 5 cycles of 95 °C for 5 s; 52 °C for 1 min; and 72 °C for 30 s, 10-12 cycles of 95 °C for 5 s; 60 °C for 30 s; and 72 °C for 30sec, and 72 °C for 2 min. The resultant amplicons were purified with Qiaquick PCR purification columns and then re-amplified by PCR in 32 reactions (100 µl each) with

0.02 nM first round amplicons, 400 nM each of eMIP_CA1_F primer and eMIP_CA1_R primer, and 50 µl of KAPA SYBR fast Universal qPCR Master Mix at 95 °C for 30 s, 13-15 cycles of 95 °C for 5 s; 60 °C for 30 sec; and 72 °C for 30 s, and 72 °C for 2 min. The resultant amplicons were purified by ethanol precipitation and re-purified with Qiaquick PCR purification columns as described above. Approximately 4 µg of the purified amplicons were digested with 100 units of Nt.AlwI (100 U/µl, NEB) at 37 °C for 1 h in NEBuffer 2. The enzyme was inactivated by incubation at 80 °C for 20 min. The digested amplicons were then incubated with 100 units of Nb.BrsDI (10 U/µl, NEB) at 65 °C for 1 h. The resultant nicked DNA was purified with Qiaquick PCR purification columns. The probe molecules (of approximately 70 bases in length) were purified by 6% denaturing PAGE (6% TB-urea 2D gel).

2.3.4 Sample preparation and capture

Genomic DNA was extracted using the AllPrep DNA/RNA Mini kit (Qiagen) and bisulfite converted with the EZ-96 DNA methylation Gold kit (Zymo Research) in a 96-well plate. Normalized amounts of padlock probes, 200 ng bisulfite converted gDNA, and 4.2 nM oligonucleotide suppressors were mixed in 25 µl 1x Ampligase Buffer (Epicentre) in a 96-well plate format and incubated at 95 °C for 10 min; the temperature was then gradually lowered at a rate of 0.02 °C/s to 55 °C and held for 20 h. 2.5 µl of SLN mix (100 µM dNTP, 2 U/µl AmpliTaq Stoffel Fragment (ABI), and 0.5 U/µl Ampligase (Epicentre) in 1 × Ampligase buffer) was added to each mixture for gap-filling and circularization; the reactions were then incubated at 55 °C for 20 and then at 94 °C for 2 min for enzyme inactivation. To digest linear DNA after circularization, 2 µl of exonuclease mix (10 U/µl exonuclease I and 100 U/µl exonuclease III (USB)) was added to each reaction; the reactions were then incubated at 37 °C for 2 h and then at 94 °C for 2 min for enzyme inactivation.

2.3.5 Capture circles amplification (Library-free protocol, Agilent)

10 µl circularized DNA was amplified and barcoded in 100 µl reactions with 400 nM each of AmpF6.3Sol primer (Supplementary Table 2.3) and AmpR6.3 indexing primer (Supplementary Table 2.3), 0.4x SYBR Green I (Invitrogen), and 50 µl Phusion High-Fidelity 2x Master Mix (NEB) at 98 °C for 30 s, 5 cycles of 98 °C for 10 s; 58 °C for 20 s; and 72 °C for 20 s, 9-12 cycles of 98 °C for 10 s; and 72 °C for 20 s, and 72 °C for 3 min.

2.3.6 Capture circles amplification (Library-free protocol, LC Sciences)

10 µl circularized DNA was amplified in a 100 µl reaction with 200 nM each of CP-2-FA primer and CP-2-RA primer (Supplementary Table 2.3) and 50 µl KAPA SYBR fast Universal qPCR Master Mix at 98 °C for 30 s, 5 cycles of 98 °C for 10 s; 52 °C for 30 s; and 72 °C for 30 s, 15 cycles of 98 °C for 10 s; 60 °C for 30 s; and 72 °C for 30 s, and 72 °C for 3 min. The resultant amplicons with the corresponding expected size of approximately 260 bp were purified with 6% PAGE (6% 5-well gel, Invitrogen) and resuspended in 12 µl of TE buffer. 30% of the gel-purified amplicons were re-amplified and barcoded in a 100 µl reaction with 200 nM each of two different sets of primers to enable SE sequencing of both ends of the amplicons (CP-2-FA.IndSol primer and CP-2-RA.Sol primer or Switch.CP-2-FA and Switch.CP-2-RA.IndSol) and 50 µl KAPA SYBR fast Universal qPCR Master Mix at 98 °C for 30 s, 4 cycles of 98 °C for 10 s; 54 °C for 30 s; and 72 °C for 30 s, and 72 °C for 3 min.

2.3.7 Primer barcode design for multiplexing

A Perl script was written to randomly generate 6 nucleotide long sequences. A sequence was kept as long as it did not have more than two matching positions with another accepted barcode and as long as it had between two to four guanine/cytosine

residues. This process was repeated until the desired number of barcodes was obtained. A total of 384 primers were designed.

2.3.8 Bisulfite read mapping and data analysis

The reference genome was computationally bisulfite converted by changing all C's to T's on Watson and Crick strands separately. FASTQ reads were encoded by first predicting the mapping orientation, and then converting all predicted forward mapping reads by changing all C's to T's and converting all predicted reverse mapping reads by changing all G's to A's. The reads were then mapped to the converted genome with SOAP2Align (http://soap.genomics.org.cn/soapaligner.html) in paired-end mode with the parameters r = 0, v = 2 (one mismatch per 40bp sequenced), m = 0, x = 400. Alignment files were then combined, and the most accurate alignment for each read was selected. Original C and G calls were then placed back into the alignment information. Alignments were then converted to pileup format using SamTools (http://samtools.sourceforge.net/). Raw SNPs and methylation frequency files were computed from pileup counts. Methylation frequencies and SNPs were called using a method described previously[11].

2.3.9 Correlation of methylation levels between two samples

To check if methylation levels were similar between two samples, the Pearson's correlation was calculated using all CpG sites characterizable in both. First, a list of CpG sites with read depth of at least 10 in both samples was generated. The methylation frequencies at these sites were obtained from bisReadMapper output, and then loaded into the statistical package R. Finally, the Pearson's correlation between the two samples was computed using the cor() function[34].

2.3.10 Analysis of differential methylation

From the bisReadMapper output, the number of reads showing methylated and unmethylated cytosine at each CpG site was calculated. Using these counts, a Fisher-Exact Test with Benjamini-Hochberg Multiple Testing Correction (FDR = 0.01) was carried out between each pair of cell lines using CpG sites covered by at least 10 reads in both lines to generate a list of statistically different sites. Cases where the absolute methylation level was within 0.1 were then removed. This resulted in a set of differentially methylated sites (DMSs) between each pair of lines. Technical replicates did not show any differential methylation, while different cell types showed a large degree (~33%).

2.4 Results

To interrogate the methylation of the most informative loci across many samples quickly and cost-effectively we developed the second generation BSPP for improved flexibility and multiplexing capability. These improvements have contributed to recent findings in mouse and human pluripotent stem cells[30-32, 35].

First, target selection and probe design is crucial for BSPP. To aid in the design of efficient padlock probes for bisulfite analysis, we developed a program called ppDesigner. It accepts as input the genome of any organism, a list of arbitrary targets desired by the user, and a set of user-desired probe constraints matching requirements of the experimental protocol. It *in silico* bisulfite-converts the genome on the fly (that is, it changes all cytosine to thymine) and outputs a set of padlock probes to cover the chosen targets while avoiding CpGs in the capturing arms, which could be methylated and not converted to be recognized as thymine. ppDesigner uses a back-propagation neural network to predict probe efficiency (Supplementary Fig. 2.1).

We had previously trained this network using data from probes for exomic targets[36] based on seven properties. Using bisulfite capture data, we have refined the network with two additional factors. ppDesigner can explain ~50% of the variance in capturing efficiency for genomic DNA and ~20% of the variance in capturing efficiency for bisulfite converted DNA; additional variation could be due to factors such as variability in oligonucleotide synthesis and sample DNA quality.

Key requirements for methylation analysis on large sample sizes include low cost, simple workflow, and automation compatibility. As DNA sequencing cost has rapidly decreased, sample processing has become a bottleneck in terms of cost and throughput. A complicated workflow increases variability between samples, and reduces power in large-scale studies. To address these issues, we extended a "library-free" protocol[37] to multiplexed BSPP capture (Fig. 2.1). This method eliminates five steps from Illumina's library construction protocol, such that multiplexed libraries can be generated from DNA in only four steps (Supplementary Table 2.1). Using multiplexed primers with 6 bp barcodes, we have routinely generated libraries for 96 samples in 96-well plates and sequenced all at once in a single Illumina HiSeq flowcell. Additional primers have been designed to process 384 samples per batch. As sample-specific barcodes were added, barcoded libraries can be pooled for size-selection, which is the most time consuming, contamination-prone, and error-prone step if performed individually. The protocol is compatible with multi-channel pipettes or liquid handling devices. It dramatically reduced experimental cost and time, and improved reproducibility and read mapping rates (Supplementary Tables 2.1 and 2.2). For large sample sizes, the library preparation cost (including probes) is comparable to that of the Restricted Representation and Whole Genome Bisulfite Sequencing (RRBS, WGBS) protocols, while the sequencing cost is much

lower than that of WGBS due to targeting of CpG sites of interest. RRBS is more cost-effective than BSPP, but there is little flexibility in selecting specific sites or regions.

Another bottleneck in bisulfite sequencing is a lack of computational tools to efficiently analyze sequencing data generated from hundreds of samples. To overcome this issue, we developed an analysis pipeline for read mapping and methylation quantification called bisReadMapper (Supplementary Fig. 2.2). In previous padlock probe studies, reads were mapped only against target regions due to the computational requirements of sequence alignment[11]. In contrast, bisReadMapper maps to the full genome sequence, allowing processing of both targeted and whole genome bisulfite data. bisReadMapper also determines the origin strand of the read based on base composition and maps reads as if they were fully bisulfite-converted to a fully bisulfite-converted genome sequence, allowing mapping of both bi- and uni-directional bisulfite libraries in an unbiased manner. Another feature is the capability to call single nucleotide polymorphisms from bisulfite sequencing data; this feature not only allows for analysis of allele-specific methylation[38], but also allows accurate sample tracking in large-scale experiments. Finally, bisReadMapper can call methylation levels at both CpG and non-CpG sites.

To demonstrate the effectiveness of our assay, we generated a new genome-scale probe set based on our previous results and new information about differential methylation[11, 39-41]. Our new design was targeted to evaluate the methylation level at a set of genomic locations known to contain differentially methylated regions (DMRs) or sites (DMSs)[39-42], CTCF binding sites, and DNase I hypersensitive regions. In addition, all microRNA genes and all promoters for human NCBI Reference Sequence (RefSeq) genes were targeted. Using ppDesigner, we successfully

designed ~330,000 padlock probes that covered 140,749 non-overlapping regions with a total size of 34 megabases. We performed capturing experiments and end-sequencing, and found that these probes were slightly more specific (~96% on-target) and uniform than previous probes[11] (Supplementary Fig. 2.3). These probes were further normalized using subsetting and suppressor oligonucleotides as described previously[11] to improve uniformity. Roughly 500,000 CpG sites were characterizable with ~4 gigabases of sequencing reads, and additional sites became callable with deeper sequencing (Supplementary Fig. 2.4 and 2.5).

We used this probe set to analyze H1 embryonic stem cells (H1 ESCs), PGP1 fibroblasts (PGP1F), and two technical replicates of PGP1 fibroblast-derived induced pluripotent stem cells (PGP1-iPSC). For each sample, we sequenced on average ~3.66 gigabases and measured the methylation level for an average of 480,904 CpG sites. In order to assess whether this data could identify potential epigenetic regulation of transcription, we utilized GREAT[43] to predict the cis-regulatory potential of regions around captured CpG sites. In total, the padlock probes captured CpG sites in regions predicted to regulate 98% of RefSeq genes (Supplementary Fig. 2.6).

The data generated by BSPP accurately represented the methylation status of the target regions. Methylation levels for the two technical replicates of PGP1-iPSC were consistent both within a single batch and between separate batches (Pearson's correlation coefficient R = 0.97 – 0.98, Supplementary Fig. 2.7a,b). Additionally, when methylation levels were compared between technical replicates, no CpG site was found to be significantly different by a Fisher Exact Test with Benjamini-Hochberg multiple testing correction (FDR = 0.01, n = 439,090). In comparison, large fractions of sites were found differentially methylated due to either the process of nuclear reprogramming (27.9% DMS between PGP1-iPSC and PGP1F) or the difference in

cell type (31.3% DMS between PGP1F and H1) with the same criteria (FDR = 0.01, n = 444,111 and 359,290, respectively). Our BSPP results on H1 ESCs are highly consistent with the published whole genome bisulfite sequencing data[41] (Pearson's correlation coefficient R = 0.95, Supplementary Fig. 2.8).

Our assay has very low technical variability. We have performed the assay on over 150 samples in 96-well plates; the yield for each was similar (Supplementary Fig. 2.9). Approximately 10% of CpG sites are targeted separately on each strand, allowing low-quality data sets with poor correlation between these built-in technical replicates to be identified (Supplementary Fig. 2.7c,d,e). As our BSPP assay measures absolute methylation levels, no normalization is necessary as long as the internal replicates are consistent. Therefore, a large number of datasets, even generated from different laboratories, can be directly compared without batch effects, which is important for case-control studies on large samples or meta-analyses. Additionally, the SNP-calling feature of bisReadMapper allowed us to characterize roughly 20,000 SNPs for each sample at an accuracy of 96% or better. This allowed us to unambiguously track samples, which is crucial for projects involving large sample sizes.

Our library-free BSPP method is flexible for different study designs. While our genome-scale probe set allows global profiling on thousands of samples, a focused assay is often necessary to follow up on tens to hundreds of candidate regions identified in genome-scale scanning. Such an assay needs to be customizable to different genomic targets, scalable to a very large sample size (1,000-100,000), and inexpensive. To further demonstrate the flexibility, we designed a second set of 3,918 probes to evaluate the methylation state 1 kb upstream and downstream of 120 genomic regions previously known and confirmed by BSPP to carry aberrant

methylation in induced pluripotent stem cells[26]. We acquired the oligonucleotides from a second vendor (LC Sciences). Even with shorter capturing sequences (40 bp total for capturing arms rather than 50 bp on average, Supplementary Fig. 2.10) and a 100-fold smaller target size, an average of 56% of mappable bases were on-target, equivalent to an enrichment factor of ~6,500x. With the data from three cell lines (H1 ESCs, PGP1F, and PGP1-iPSCs) we were able to identify regions of aberrant methylation in iPSCs (Supplementary Fig. 2.11), and demonstrated that aberrant methylation continues further upstream and downstream than observed previously. This analysis demonstrates that a focused probe set can be used to validate specific regions of interest identified in global scanning using either our genome-wide probe set or other methods.

This method can be implemented to aid in identifying the effects of DNA methylation in any organism by using the computational tools made available on the supporting website for this paper (http://genome-tech.ucsd.edu/public/Gen2_BSPP/).

2.5 Conclusions

We have demonstrated that we can utilize padlock probes to quickly and easily perform targeted capture and sequencing on a wide variety of interesting genomic regions. While this chapter has primarily discussed the use of padlock probes to analyze DNA methylation, probes can also be used to target genomic features of interest; we have previously utilized padlock probes to capture and analyze both SNPs (with 132,000 probes) and subsets of exomic targets (with 220,000 probes)[10]. Due to their high efficiency, low cost, and high degree of scalability, padlock probes are a valuable tool for high-throughput DNA sequencing experiments in both a genomic and an epigenomic context.

2.6 Acknowledgements

We thank A. Feinberg and J. Stamatoyannopoulos for providing informative genomic targets and E. LeProust (Agilent Technology) for long-oligonucleotide synthesis. This work is funded by grants from National Institute of Health (R01 DA025779; R01 GM097253) to Kun Zhang.

Chapter 2, in part, is a reprint of the material as it appears in: Dinh Diep*, Nongluk Plongthongkum*, Athurva Gore*, Ho-Lim Fung, Robert Shoemaker, Kun Zhang. “Library-free Methylation Sequencing with Bisulfite Padlock Probes.” *Nature Methods*. 2012 February 5; 9(3): 270-272. doi: 10.1038/nmeth.1871. Used with permission. The dissertation author was one of the primary investigators and authors of this paper.

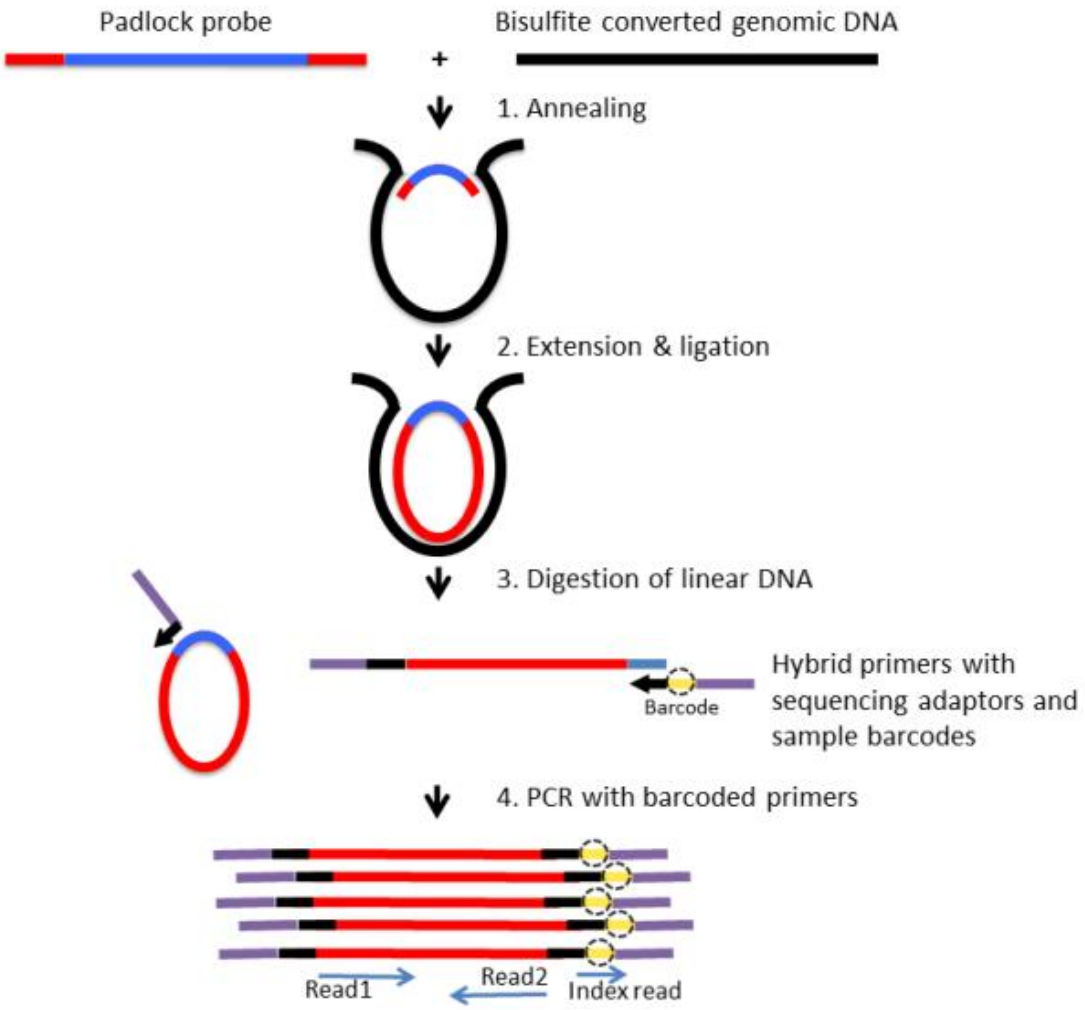
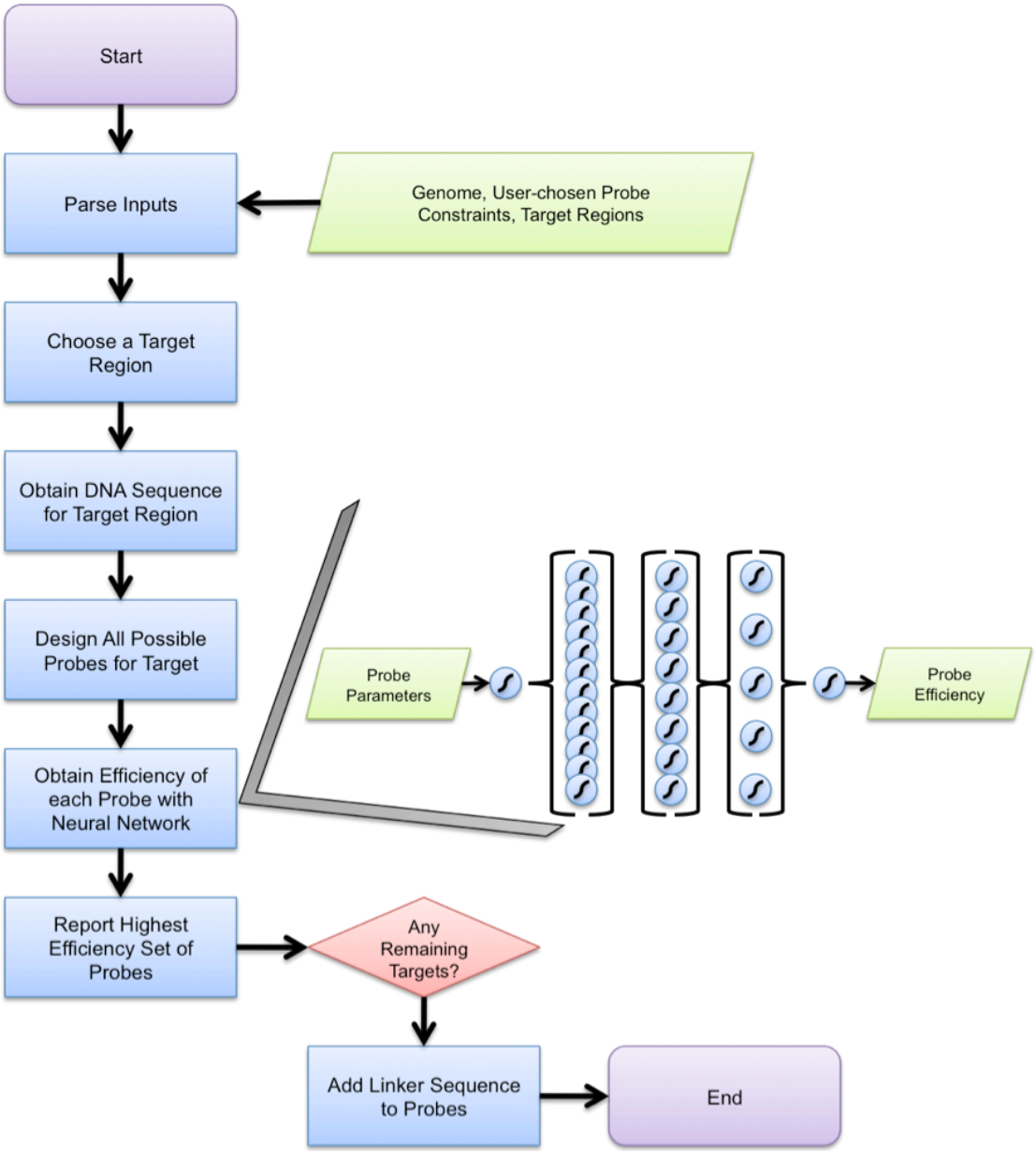
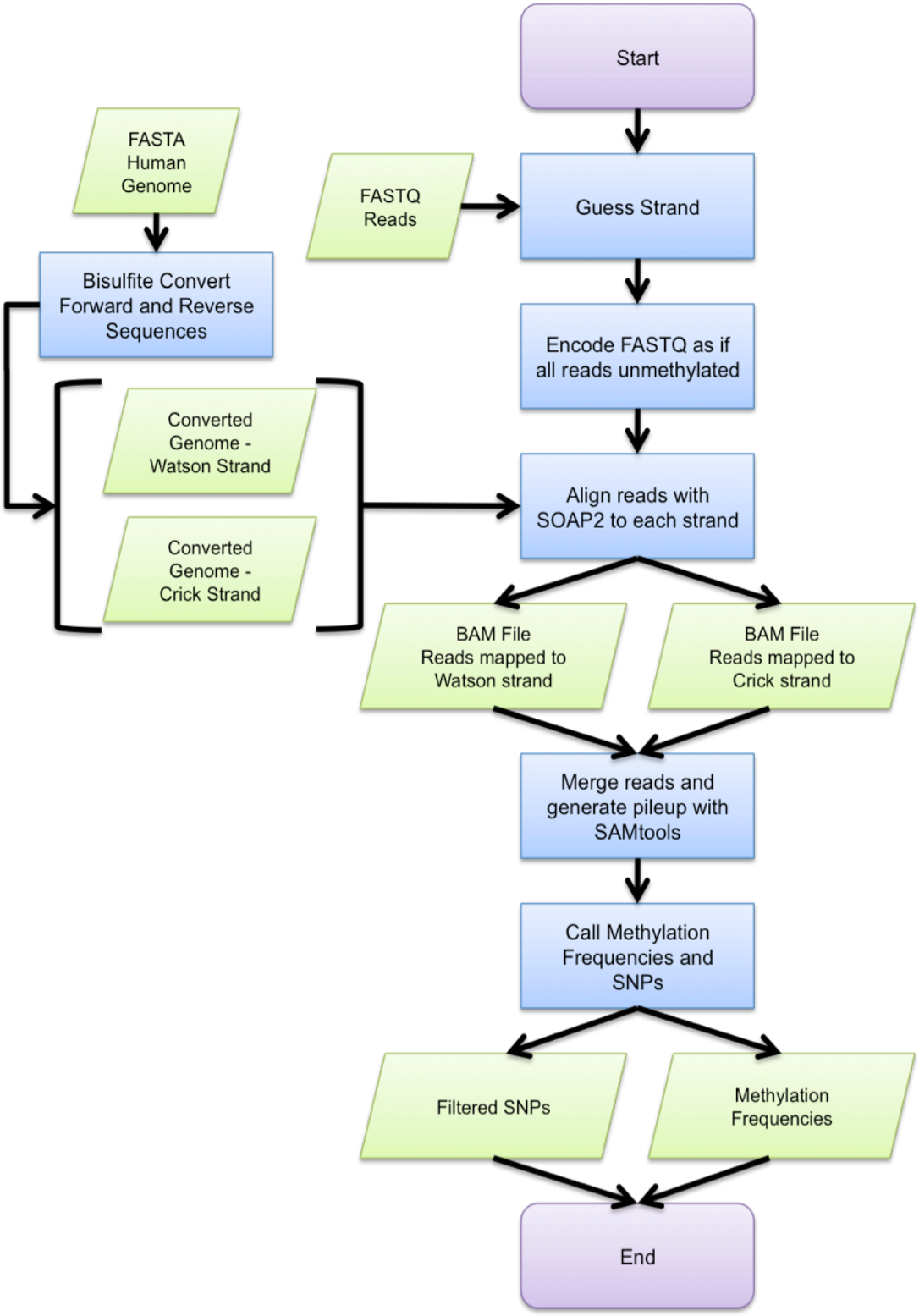


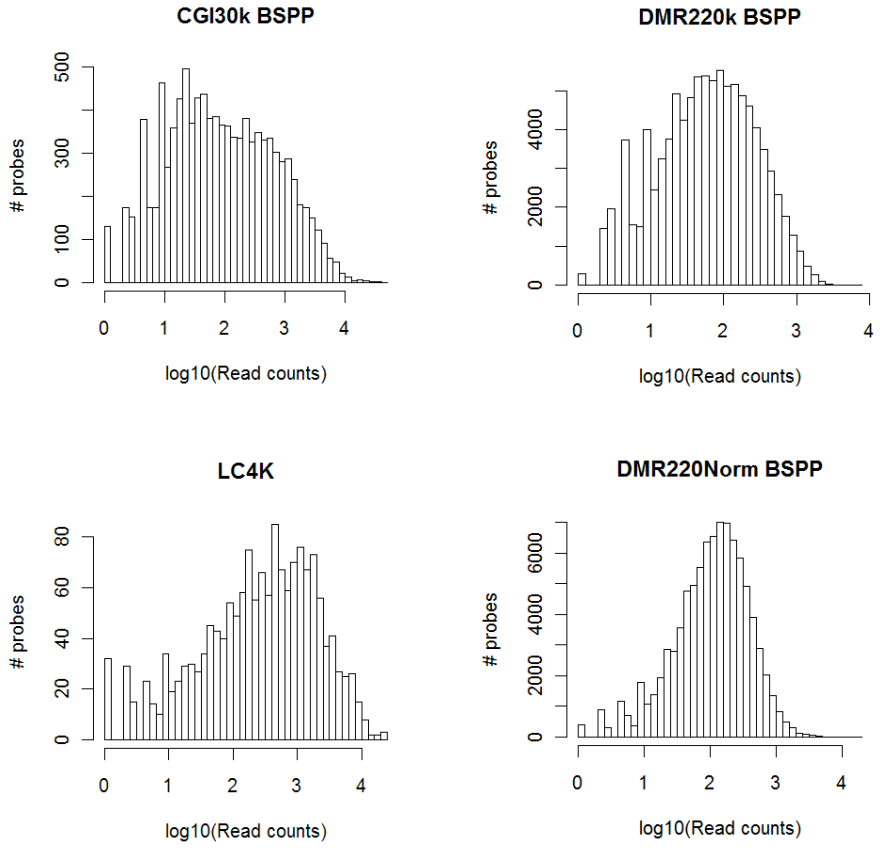
Figure 2.1. Schematic of library-free BSPP protocol. Each padlock probe has a common linker sequence flanked by two target-specific capturing arms (red) that anneal to bisulfite converted genomic DNA (black). The 3′ end is extended and ligated with the 5′ end to form circularized DNA. After removal of linear DNA, all circularized captured targets are PCR-amplified with barcoded primers and directly sequenced with an Illumina sequencing platform (GA II(x) or HiSeq). Amplicon size is 363 bp, which includes captured target (180 bp), capturing arms (55 bp), and amplification primers and adapters (128 bp). The inserts can be read through with paired-end 120 bp sequencing reads.



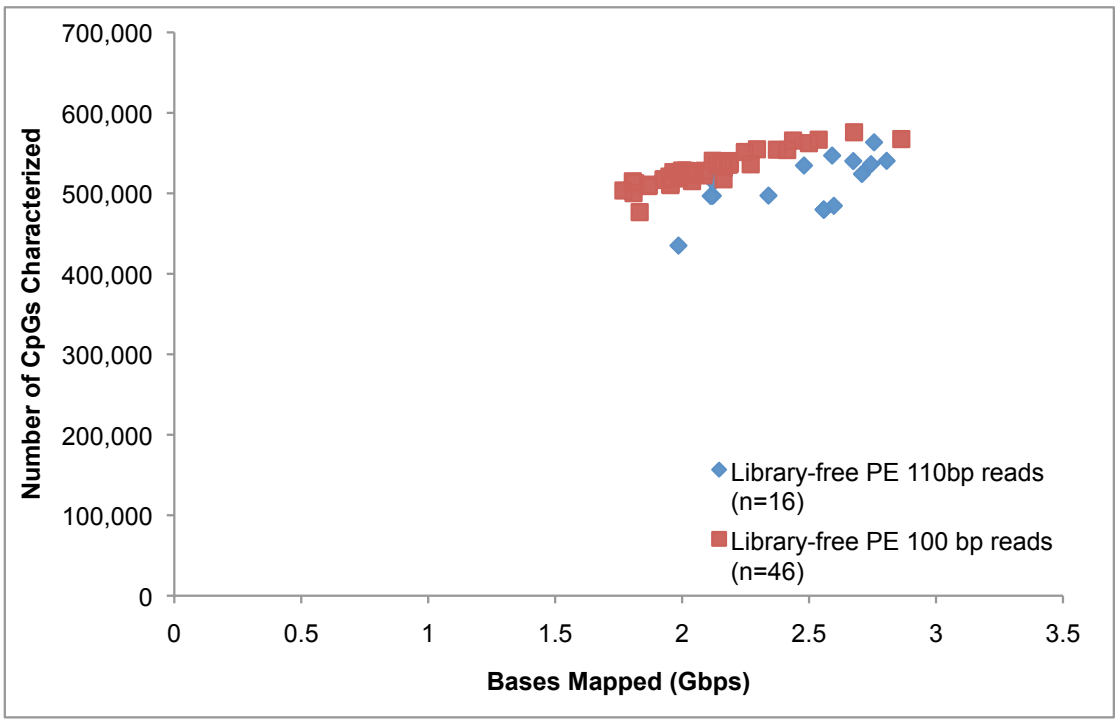
Supplementary Figure 2.1. Schematic for the probe design software (ppDesigner). The neural network model utilizes the target length, target GC content, binding arm melting temperature, binding arm length, local single-stranded folding energy of the target, and the dinucleotides present at the extension site and ligation site during probe capture. Example probes can be found in Supplementary Figure 10.



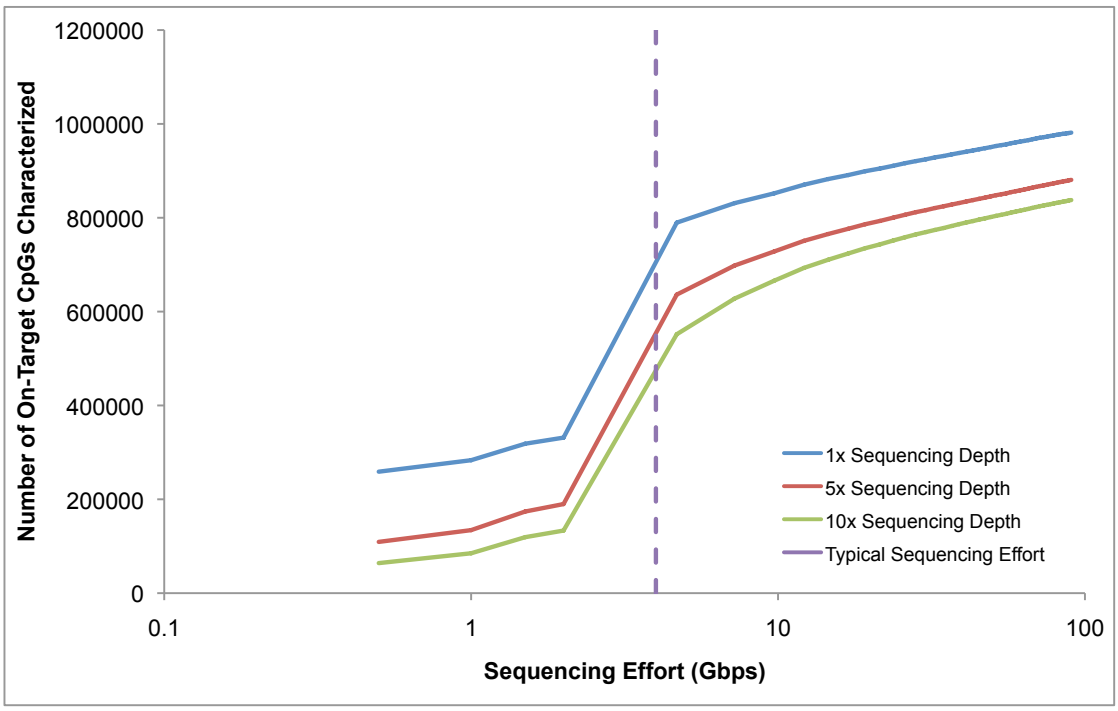
Supplementary Figure 2.2. Schematic for the bisulfite sequencing data analysis pipeline (bisReadMapper).



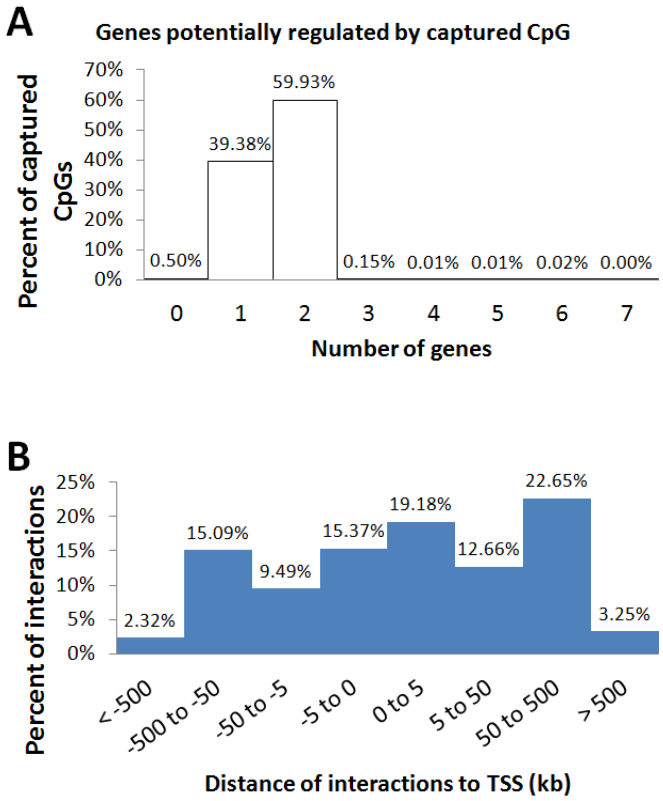
Supplementary Figure 2.3. Comparison of probe capture efficiencies between the DMR220K, LC4K probe sets and the previously published CGI30K set. The first three plots were generated from data without subsetting or suppressor oligonucleotides to allow for a direct comparison of probe design.



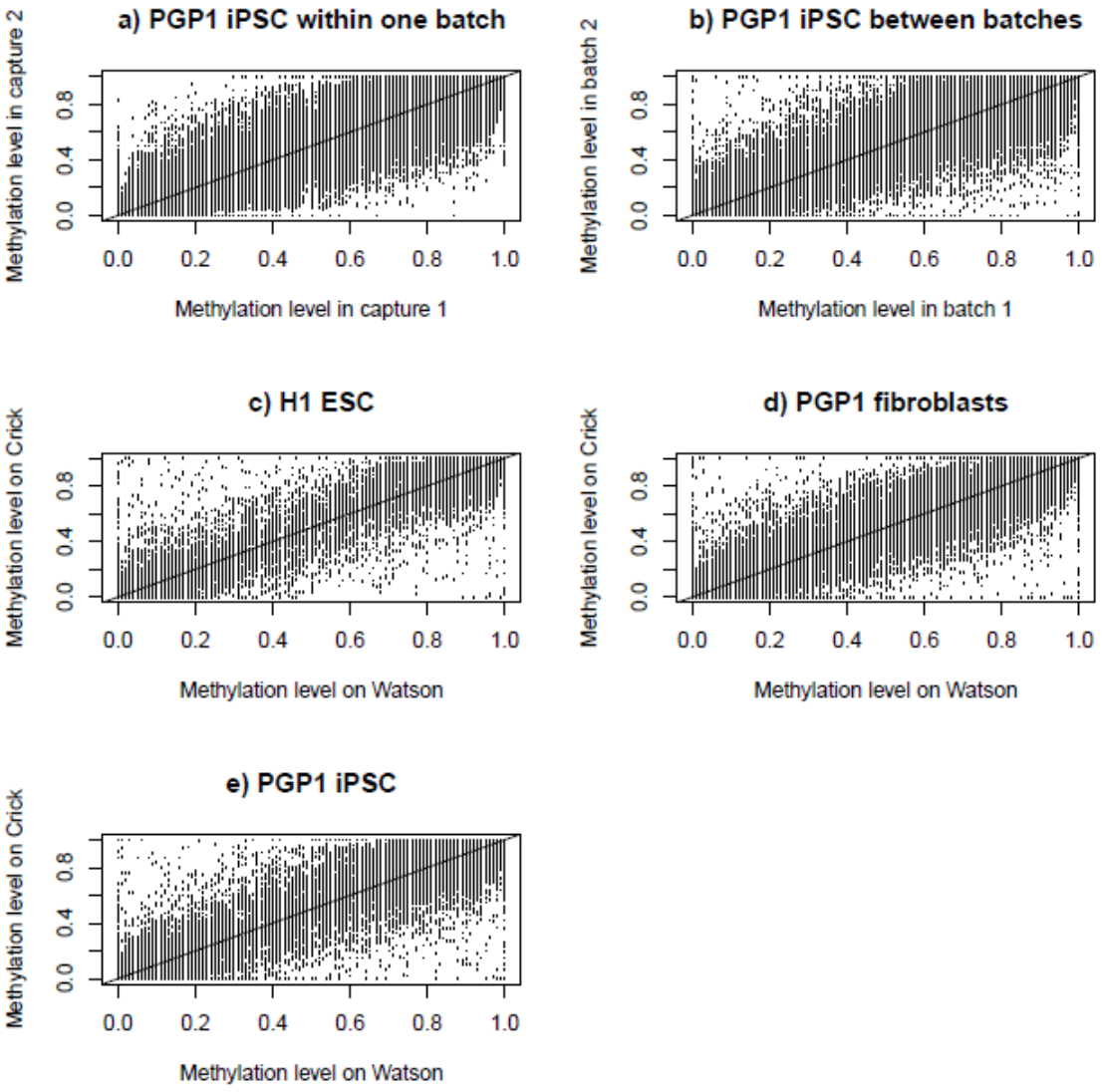
Supplementary Figure 2.4. Scatter plot of number of characterized CpG sites versus mappable sequencing data for the DMR330K probe set. Variability in sequencing quality of individual sequencing runs is responsible for the different number of CpG sites characterized with similar sequencing effort.



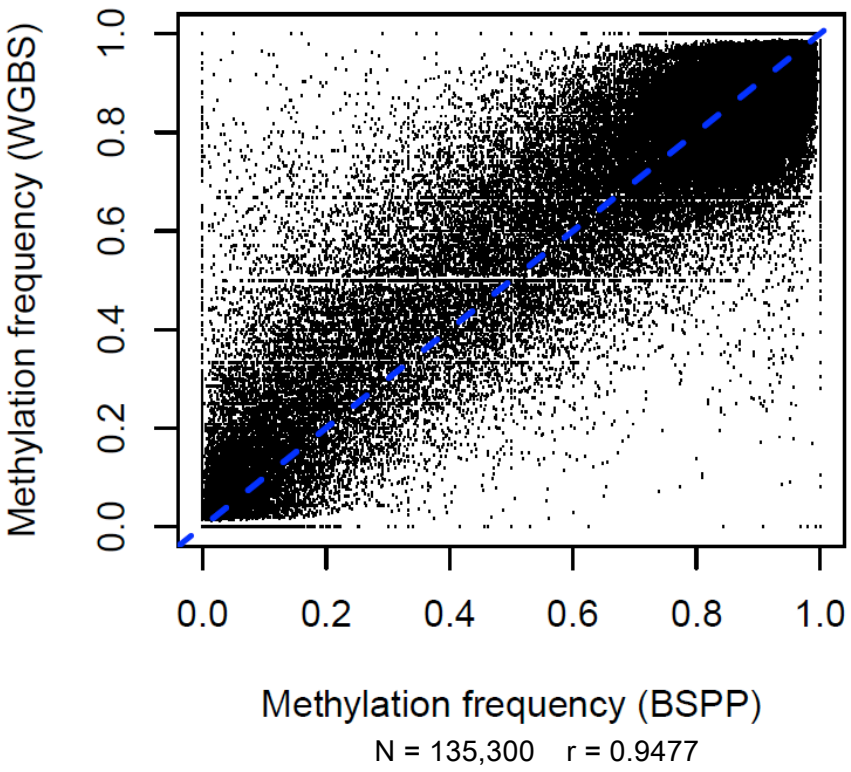
Supplementary Figure 2.5. Number CpG sites called per sample as a function of sequencing effort. The horizontal dash line represents 4Gbps of sequences per library that we routinely generate.



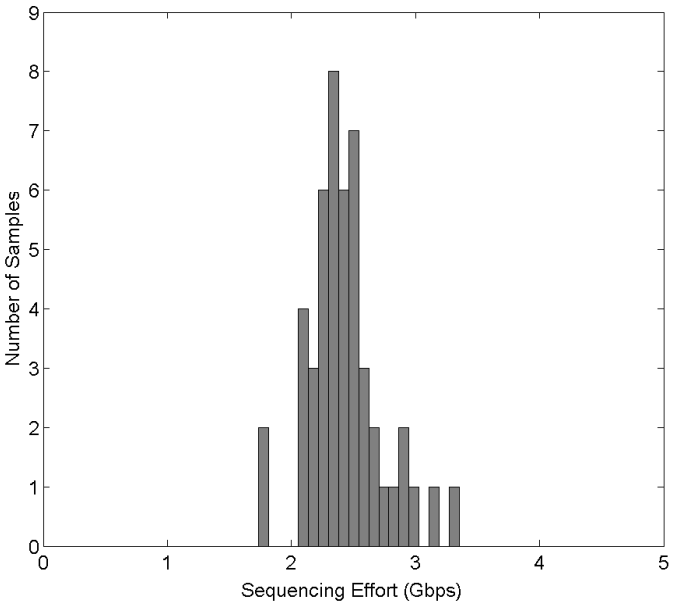
Supplementary Figure 2.6. Captured CpG sites were tested for potential regulatory interactions with genes by GREAT (http://great.stanford.edu). (A) Most CpG sites were interacting with 1-2 genes. (B) Distance of CpG sites to the transcriptional start sites (TSS) of the predicted regulating genes.



Supplementary Figure 2.7. Accuracy of digital quantification by BSPP. (a,b) Within batch and between batch comparison of the methylation levels obtained at 10x depth from multiple capture reactions of the same sample (PGP1iPS). The Pearson's correlation coefficient R for within one batch is 0.98 (N=405,508), and for between batches is 0.97 (N=117,186). (c,d,e) Within sample comparison of methylation levels obtained from different probes capturing the same CpG site on different strands at 10x depth within one capture reaction. The Pearson's correlation coefficient R was 0.96 (N=44,361), 0.96 (N=55,965), and 0.97 (N=29,884) for PGP1iPS, PGP1F, and H1 respectively.

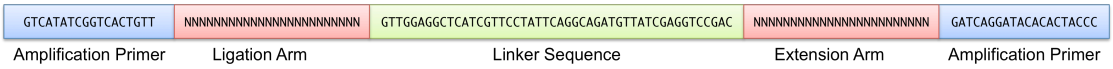


Supplementary Figure 2.8. Comparison between BSPP and whole genome bisulfite sequencing (WGBS). We compared two H1 ESC datasets, using sites with at least 10x read depth in each. The Pearson's correlation coefficient R was 0.9477 (N=135,300). (Note that the sequencing experiments were performed on separate cultures of H1 from two different labs.)

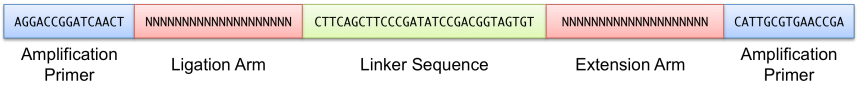


Supplementary Figure 2.9. Variation in amount of sequencing data obtained per sample in a multiplexed BSPP capture experiment. 48 whole blood samples were captured and sequenced in one batch using the library-free BSPP method. There is little variation between samples in the amount of generated sequencing data.

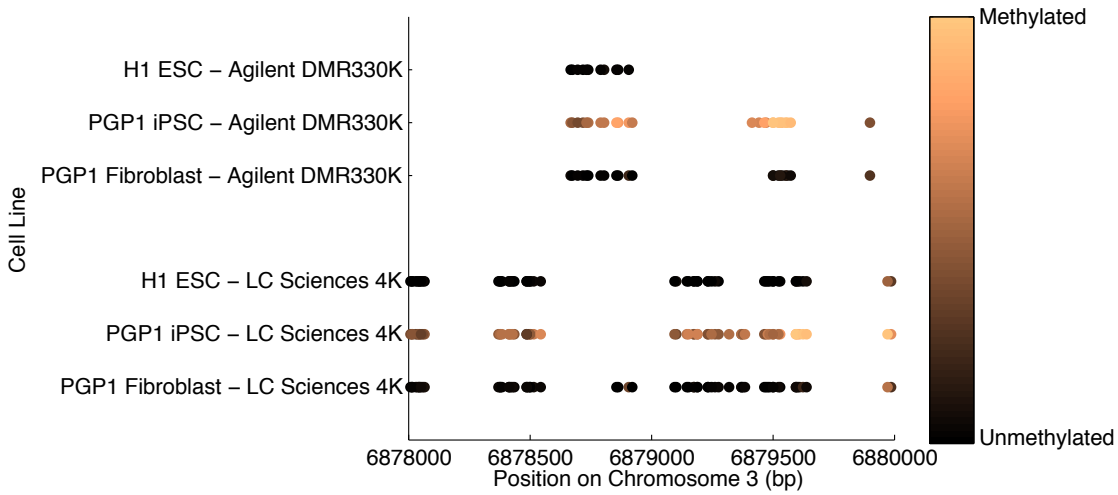
A. Agilent Padlock Probe



B. LC Sciences Padlock Probe



Supplementary Figure 2.10. Example padlock probes ordered from (A) Agilent's oligonucleotide synthesis service and (B) LC Sciences' oligonucleotide synthesis service.



Supplementary Figure 2.11. UCSC Genome Browser view showing an example of aberrant iPSC-specific methylation after reprogramming of PGP1 fibroblasts into iPS cells. Circles represent a location with measurable methylation state, with black indicating unmethylated and gold indicating methylated. The Agilent 330K probe set identified a small intronic region containing aberrant methylation in the iPS cells that are not present in either the fibroblast progenitors or a control hESC line. The LC Sciences 4K probe set was designed to characterize the methylation state upstream and downstream of this region. This focused assay revealed that the abnormal methylation also extended into the exomic region of GRM7.

Supplementary Table 2.1. Comparison of bisulfite sequencing methods. The number of enzymatic reactions, number of purifications, cost per sample, and mapping rates for first-generation padlock probes, second-generation library-free padlock probes, reduced representation bisulfite sequencing (RRBS), and whole genome bisulfite sequencing (WGBS) are shown.

	Published BSPP	Library-free BSPP	RRBS	WGBS
Enzymatic reactions	10	3	4	3
Purification	6	1	3	3
Size-selection	2	1[1]	1	1
Sample preparation cost per sample	$71.15[1]	$37.86[2]	$28.15	$31.10
Mapping rate	44%	87%	27%[3]	N.D.
Genome coverage obtained at 10x depth	<0.1%	0.6%-1%	~1%[3]	76-96%[4]
Sequencing (Gbps)	0.5	4.0	1.4	70.0
Sequencing cost per sample[5]	$24.38	$195.00	$68.25	$3412.50

1 Unlike other methods, in the library-free BSPP protocol size selection is typically performed on 48-96 pooled libraries.

2 Includes the cost of ordering 400,000 synthesized probes from LC Sciences and reagents for preparing probes, bisulfite conversion, capture, and sequencing library preparation. Estimates assume that 10,000 samples will be processed.

3 Estimated from: Gu et. al., Nat Methods 2010; 7(2):133-136.

4 Adapted from: Beck et. al., Nat Biotechnol 2010;28:1026-1028.

5 Assumes sequencing using an Illumina HiSeq to generate 300 Gbps of sequencing data, with cost of $4920 for a flowcell, $6815 for sequencing reagents, and $2890 for service fee. ($48.75 per Gbps)

Supplementary Table 2.2. Representative cost per sample for oligonucleotide synthesis, sequencing library construction, and Illumina sequencing.

Expected number of samples to be processed	Probe set sizes		
	4,000	40,000	400,000
10	$134.57	$872.04	$9,298.78
100	$35.57	$129.54	$1,131.28
1000	$25.67	$55.29	$314.53
10000	$24.68	$47.86	$232.86

Supplementary Table 2.3. Primer sequences used for padlock probe production, padlock capture, sequencing library construction, and Illumina sequencing.

Primer name	Primer sequences
Primers used with Agilent Probes	
pAP1V61U	5'-G*G*G*TCATATCGGTCACTGTU-3'
AP2V6	5'-/5Phos/CACGGGTAGTGTGTATCCTG-3'
RE-DpnII-V6	5'-GTGTATCCTGATC-3'
AmpF6.4Sol	5'-AATGATACGGCGACCACCGAGATCTACACCACTCTCAGATGTTATCGAGGTCCGAC-3'
AmpF6.3NH2	5'-/5AmMC6/CAGATGTTATCGAGGTCCGAC-3'
AmpR6.3NH2	5'-/5AmMC6/GGAACGATGAGCCTCCAAC-3'
PCR_F	5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTC-3'
PE_t_N2	5'-ACACTCTTTCCCT ACACGACGCTCTTCCGA TCTN*N-3'
PE_b_A	5'-/5Phos/AGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'
SolSeq6.3.3 (Read1)	5'-TACACCACTCTCAGATGTTATCGAGGTCCGAC -3'
SolSeqV6.3.2r(Read2)	5'-GCTAGGAACGATGAGCCTCCAAC-3'
AmpR6.3IndSeq(IndexRead)	5'-GTTGGAGGCTCATCGTTCCTAGC-3'
Primers used with LC Sciences Probes	
eMIP_CA1_F	5'- TGCCTAGGACCGGATCAACT-3'
eMIP_CA1_R	5'- GAGCTTCGGTTCACGCAATG-3'
CP-2-FA	5'-GCACGATCCGACGGTAGTGT-3'
CP-2-RA	5'-CCGTAATCGGGAAGCTGAAG-3'
CA-2-FA.Indx7Sol	5'-CAAGCAGAAGACGGCATACGAGATGATCTGCGGTCTGCCATCCGACGGTAGTGT-3'
CA-2-FA.Indx45Sol	5'-CAAGCAGAAGACGGCATACGAGATCGTAGTCGGTCTGCCATCCGACGGTAGTGT-3'
CA-2-FA.Indx76Sol	5'-CAAGCAGAAGACGGCATACGAGATAATAGGCGGTCTGCCATCCGACGGTAGTGT-3'
CA-2-RA.Sol	5'- AATGATACGGCGACCACCGAGATCTACACGCCTATCGGGAAGCTGAAG-3'
Switch.CA-2-FA.Sol	5'- AATGATACGGCGACCACCGAGATCTACACGCCTATCCGACGGTAGTGT-3'
Switch.CA-2-RA.Ind7Sol	5'-CAAGCAGAAGACGGCATACGAGATGATCTGCGGTCTGCCATCGGGAAGCTGAAG-3'
Switch.CA-2-RA.Ind45Sol	5'-CAAGCAGAAGACGGCATACGAGATCGTAGTCGGTCTGCCATCGGGAAGCTGAAG-3'
Switch.CA-2-RA.Ind76Sol	5'-CAAGCAGAAGACGGCATACGAGATAATAGGCGGTCTGCCATCGGGAAGCTGAAG-3'
CP-2-SeqRead1.x (Read1)	5'-TACACGCCTATCGGGAAGCTGAAG-3'
CP-2-IndSeq.x (IndexRead)	5'-ACACTACCGTCGGATGGCAGACCG-3'
CP-2-SeqRead1.y (Read1)	5'-TACACGCCTATCCGACGGTAGTGT-3'
CP-2-IndSeq.y (IndexRead)	5'-CTTCAGCTTCCCGATGGCAGACCG-3'

* Indicates a phosphorothioate bond

Chapter 3: Identification of Coding Mutations in Induced Pluripotent Stem Cells

3.1 Abstract

Defined transcription factors can induce epigenetic reprogramming of adult mammalian cells into induced pluripotent stem cells. Although DNA factors are integrated during some reprogramming methods, it is unknown whether the genome remains unchanged at the single nucleotide level. Here we show that 22 human induced pluripotent stem (hiPS) cell lines reprogrammed using five different methods each contained an average of five protein-coding point mutations in the regions sampled (an estimated six protein-coding point mutations per exome). The majority of these mutations were non-synonymous, nonsense or splice variants, and were enriched in genes mutated or having causative effects in cancers. At least half of these reprogramming-associated mutations pre-existed in fibroblast progenitors at low frequencies, whereas the rest occurred during or after reprogramming. Thus, hiPS cells acquire genetic modifications in addition to epigenetic modifications. Extensive genetic screening should become a standard procedure to ensure hiPS cell safety before clinical use.

3.2 Introduction

Human induced pluripotent stem cells have the potential to revolutionize personalized medicine by allowing immunocompatible stem cell therapies to be developed[23, 44]. However, questions remain about hiPS cell safety. For clinical use,

hiPS cell lines must be reprogrammed from cultured adult cells, and could carry a mutational load due to normal in vivo somatic mutation. Furthermore, many hiPS cell reprogramming methods use oncogenes that may increase the mutation rate. Additionally, some hiPS cell lines have been observed to contain large-scale genomic rearrangements and abnormal karyotypes after reprogramming[45]. Recent studies also revealed that tumor suppressor genes, including those involved in DNA damage response, have an inhibitory effect on nuclear reprogramming[46-51]. These findings suggest that the process of reprogramming could lead to an elevated mutational load in hiPS cells.

To probe this issue, we sequenced the majority of the protein-coding exons (exomes) of 22 hiPS cell lines and the nine matched fibroblast lines from which they came (Table 3.1). These lines were reprogrammed in seven laboratories using three integrating methods (four-factor retroviral, four-factor lentiviral and three-factor retroviral) and two non-integrating methods (episomal vector and messenger RNA delivery into fibroblasts). All hiPS cell lines were extensively characterized for pluripotency and had normal karyotypes before DNA extraction. Protein-coding regions in the genome were captured and sequenced from the genomic DNA of hiPS cell lines and their matched progenitor fibroblast lines using either padlock probes[10, 11] or in-solution DNA or RNA baits[12, 52]. We searched for single base changes, small insertions/deletions and alternative splicing variants, and identified 12,000–18,000 known and novel variants for each cell line that had sufficient coverage and consensus quality (Table 2.1).

3.3 Methods

3.3.1 CV fibroblast derivation

Primary fibroblasts were established from a 4-mm dermal punch biopsy of a 63-year-old male using a protocol based on Takashima's method[53]. The biopsy and subsequent reprogramming protocols and the informed-consent documents were reviewed and approved by the UCSD institutional ESCRO and IRB. Briefly, collagenase type 1A (Sigma) was used to dissociate the biopsy and cells were cultured in fibroblast media (DMEM containing 15% FBS, penicillin/streptomycin, sodium pyruvate, non-essential amino acids and l-glutamine). Fibroblasts were reprogrammed at passage 5. DNA was isolated for sequencing from 3,000,000 fibroblasts at passage 9.

3.3.2 CV-hiPS-B and CV-hiPS-F derivation

For reprogramming, ~100,000 fibroblasts per well were transduced with pCX4 retroviral vectors encoding OCT4 (POU5F1), SOX2, KLF4, c-MYC (MYC) and ±EGFP. CV-hiPS-B and CV-hiPS-F were derived from the +EGFP and −EGFP transductions, respectively. Transduced fibroblasts were trypsinized and seeded onto irradiated mouse embryonic fibroblasts (MEFs) and cultured in HUES media[54]. Cultures were treated with 2 mM valproic acid for the first seven days post-transduction and 10 nM Y-27632 for the first three weeks (both from EMD Chemicals). After about three weeks post-transduction, individual colonies that morphologically resembled hESCs were isolated and expanded. Established hiPS cell lines were maintained in HUES media and dissociated cultures for subculturing using 0.05% trypsin/EDTA. DNA for sequencing was isolated from CV-hiPS-B and CV-hiPS-F at passages 13 and 9, respectively.

3.3.4 CV-hiPS characterization

For PCR analysis with reverse transcription, hiPS cells were purified away from MEFs by passage onto Matrigel. Cells were collected and total RNA was isolated with the Ambion PaRIS kit following manufacturer's protocols. First-strand complementary DNA was generated with Superscript II (Invitrogen) following manufacturer's protocols. cDNA was amplified with primers specific for endogenous SOX2, NANOG and OCT4 for 30 cycles. For immunofluorescence experiments, cells were passaged onto Matrigel-coated coverslips and samples were processed using standard methods. Antibodies were used at the following dilutions: NANOG (Santa-Cruz Biotechnology, 1:200), Tra-1-81 (BD Biosciences, 1:500), SOX2 (Chemicon, 1:2,000). For embryoid body generation, hiPS cells were passaged with dispase and plated in suspension culture in embryoid body media (DMEM, 20% FBS, l-glutamine and NEAA) for eight days. On day eight, embryoid bodies were plated onto either Matrigel- or polyornithine/laminin-coated coverslips and cultured in either embryoid body media (for endoderm/mesoderm) or neural differentiation media (DMEM-F12, glutamax, N2 and B27) supplemented with dbcAMP, BDNF and GDNF (for neuroectoderm) for eight days. On day nine, cells were fixed and processed for immunofluorescence as described above. Cell Line Genetics performed karyotype analysis of CV-hiPS cell lines.

CV-hiPS-B was purified away from MEFs by culturing on Matrigel (BD Biosciences) for two passages. CV-hiPS-F was purified by dissociation with Accutase (Innovative Cell Technologies), staining with TRA-1-81 antibody (BD Biosciences) and purifying 5,000,000 TRA-1-81+ cells using a BD Biosciences FACSAria II flow cytometer.

3.3.5 dH1F-iPS8 and dH1F-iPS9 derivation

The dH1F fibroblast line was derived from the H1-OGN line previously[55]. dH1F-iPS8 and dH1F-iPS9 were reprogrammed[56] with human OCT4, SOX2, KLF4 and c-MYC retroviral vectors from dH1F at passage 5. Briefly, 293T cells in 15-cm plates were transfected with 6.25 μg of retroviral vector, 0.75 μg of VSVG vector and 5.625 μg of Gag-Pol vector using FUGENE 6 reagents. Three days after transfection, supernatants were filtered through a 0.45-μm cellulose acetate filter, concentrated by centrifugation at 23,000 r.p.m. for 90 min and stored at −80 °C until use. Transductions were carried out on dH1F fibroblast cells in six-well plates (100,000 cells per well). Viruses were added at a multiplicity of infection of five. Three days after infection, cells were split into plates pre-seeded with MEFs. The medium was changed to human ES culture medium five days after infection. hiPS cell clones stared to emerge about two to three weeks later and were picked and expanded in standard human ES cell culture medium (DMEM/F12 containing 20% KOSR, 10 ng/ml human recombinant basic fibroblast growth factor, 1x NEAA, 5.5 mM 2-ME, 50 units/ml penicillin and 50 μg/ml streptomycin). During cell collection, MEFs were removed by suction pump and collagenase (Gibco) was used to lift the cells. For dH1F, cells were cultured in 10% FBS DMEM. Trypsin-EDTA was used to lift the cells from the plate for collection. DNA was extracted using a Qiagen DNeasy kit at the following passage numbers: 12 (dH1F), 19 (dH1F-iPS8), 17 (dH1F-iPS9).

3.3.6 hiPS 11a, 11b, 17a, 17b, 29A and 29e derivation

Human fibroblasts were generated from 3-mm forearm dermal biopsies following informed consent under an IRB approved by Harvard University. The murine leukemia retroviral vector pMXs containing the human cDNAs for KLF4, SOX2 and

OCT4[57] were modified to produce higher-titer virus by including the woodchuck post-transcriptional responsive element FUGW (Addgene plasmid 14883) downstream of the cDNA. VSV-g pseudotyped viruses were packaged and concentrated by the Harvard Gene Therapy Initiative at Harvard Medical School. To produce hiPS cells, 30,000 human fibroblasts were transduced at a multiplicity of infection of 10–15 with viruses containing all three genes in hES medium with 8 μg/ml polyprene. Cells were incubated with virus for 24 h before medium was changed to standard fibroblast medium for 48 h. Cells were subsequently cultured in standard hES medium and hiPS cell colonies were manually picked on the basis of morphology within 2–4 weeks. Derived hiPS cell lines (11a, 11b, 17a, 17b and 29e) have been extensively characterized by standard assays including staining for markers of pluripotency by immunocytochemistry, cell cycle analysis, three-germ-layer differentiation potential in vitro and in vivo, and karyotype analysis[58]. All cell cultures were maintained at 37 °C in 5% CO2. Human fibroblasts were cultured in KO-DMEM (Invitrogen), supplemented with 20% Earl's salts 199 (Gibco) and 10% hyclone (Gibco), 1x GlutaMax, penicillin/streptomycin (Invitrogen) and 100 μM 2-mercaptoethanol. hiPS cells were maintained on gelatinized tissue culture plastic on a monolayer of irradiated CF-1 MEFs (GlobalStem), in hES media[54], supplemented with 20 ng/ml of bFGF. The medium was changed every 24 h and lines were passaged by trypsinization (0.5% trypsin EDTA, Invitrogen) or dispase (Gibco, 1mg/ml in hES media for 30 min at 37 °C). hiPS cell lines 11a, 11b, 17a, 17b, 29A and 29e were purified from MEFs by using dispase, which selectively detaches stem cells, and then were washed twice to ensure removal of any contaminating MEFs. Genomic DNA was extracted with a Qiagen DNeasy kit at the following passages: 7 (hFib17), 20

(iPS17A), 23 (iPS17B), 7 (hFib11), 24 (hFib11a), 20, (hFib11b), 8 (hFib29), 21 (hFib29e), 36 (hFib29A).

3.3.7 HFFxF fibroblast derivation

Primary fibroblasts were established from a foreskin biopsy of a three-year-old individual as detailed previously[59]. Briefly, a skin sample was placed in sterile saline solution, divided into small pieces and allowed to be attached to cell culture dishes before the addition of xeno-free human foreskin fibroblast growth medium. Fibroblasts generated under xeno-free conditions (HFFxF) were reprogrammed at passage 3. DNA was isolated for sequencing from 4,000,000 HFFxF fibroblasts at passage 4 with a Qiagen DNeasy kit.

3.3.8 FiPS3F1 and FiPS4F7 generation

For reprogramming, about 100,000 fibroblasts per six-well plate were transduced with 1 ml of retroviral supernatants encoding FLAG-tagged OCT4, SOX2, KLF4, and c-MYC(T58A) as described previously[60]. High-titer VSV-G-pseudotyped retroviruses expressing a polycistronic vector encoding for OCT4, SOX2, KLF4 and GFP (pMXs OSKG) and containing 5 mg/ml polybrene were produced as described[60]. Infection was performed as indicated previously[59]. Colonies were picked on the basis of morphology 25–35 days after the initial infection and plated onto fresh irradiated XF HFF (iXF HFF) cells. Xeno-free iPS cell lines FiPS3F1 and FiPS4F7 were maintained by mechanical dissociation in XF-hESm, which is composed of KO-DMEM (Dulbecco's modified Eagle's medium; Invitrogen) supplemented with 15% xeno-free KO-SR (Invitrogen), xeno-free KO-SR growth factor cocktail (1x), 2 mM glutamax, 50 mM 2-mercaptoethanol, penicillin/streptomycin (0.5x, all from Invitrogen), non-essential amino acids (Cambrex) and 20 ng/ml bFGF (Peprotech).

3.3.9 FiPS3F1 and FiPS4F7 characterization

Derived hiPS cell lines FiPS3F1 and FiPS4F7 have been extensively characterized by staining for markers of pluripotency by immunofluorescence analyses. The following antibodies were used: MAB4360 for Tra-1-60 (1:200), MAB4381 for Tra-1-81 (1:200) and AB5603 for SOX2 (1:500, all from Chemicon); MC-813-70 for SSEA-4 (1:2) and MC-631 for SSEA-3 (1:2, both from the Developmental Studies Hybridoma Bank at the University of Iowa); C-10 for OCT4 (1:100, Santa Cruz); EB06860 for NANOG (1:100, Everest Biotechnology); and Anti-FLAG (Sigma M2). Three-germ-layer differentiation potential *in vitro* was conducted by means of embryoid body formation, which was induced from colony fragments mechanically collected. For endoderm, embryoid bodies were cultured in KO-DMEM medium supplemented with 10% FBS, 2 mM l-glutamine, 0.1 mM 2-β-mercaptoethanol, non-essential amino acids and penicillin/streptomycin. For mesoderm differentiation, the same medium described above in the presence of ascorbic acid (0.5 mM) was used. For ectoderm induction, embryoid bodies were cultured in N2/B27 medium with the stromal cell line PA6 for two weeks. The medium for each condition was changed every other day. On day 15, cells were fixed and processed for immunofluorescence for the following antibodies: Tuj1 (1:500, Covance), α-fetoprotein (1:400), α-actinin (1:100, Sigma). Teratoma formation assay was performed by injecting about 0.5x 106 XF-iPS cells into the testes of severe combined immunodeficient beige mice (Charles River Laboratories). Mice were euthanized eight weeks after cell injection, and tumors were processed and analyzed following conventional immunohistochemistry protocols (Masson's trichromic stain) and immunofluorescence staining for Tuj1 (1:500, Covance), α-fetoprotein (1:400), and α-actinin (1:100, Sigma). Expression of retroviral transgenes and endogenous

pluripotency-associated factors by quantitative PCR with reverse transcription were conducted as described previously[59]. hiPS cell lines FiPS3F1 and FiPS4F7 were purified from iXF HFF by mechanical dissociation and further culturing on Matrigel (BD Biosciences) for two more passages. DNA for sequencing was isolated from passage 9 for both FiPS3F1 and FiPS4F7 with a Qiagen DNeasy kit.

3.3.10 CF-Fib, CF-RiPS1.4 and CF-RiPS1.9 derivation

CF fibroblasts (CF-Fib) were previously obtained from a skin biopsy taken from an adult with cystic fibrosis, with proper informed consent[61]. CF-induced pluripotent stem cell lines were derived using modified mRNAs coding reprogramming factors OCT4, SOX2, KLF4, c-MYC and LIN28 (OSKML) with molar concentrations in the ratio 3:1:1:1:1, in an atmosphere with 5% oxygen, as previously described[61]. Briefly, 50,000 fibroblasts were plated onto γ-irradiated human neonatal fibroblast feeders (GlobalStem) seeded at 33,00 cells/cm^2. For CF-RiPS derivations, the cationic lipid delivery system RNAiMAX was used. First, pooled RNA from the five factors OSKML (100 ng/ml) was diluted 5x and the reagent (5 µl of RNAiMAX per microgram of RNA) was diluted 10x in Opti-MEM basal media (Invitrogen). These components were pooled and incubated for 15 min at room temperature before being dispensed to culture media. Nutristem medium was replaced daily 4 h after transfection, and supplemented with 100 ng/ml bFGF and 200 ng/ml B18R (eBioscience). CF-RiPS derivation was performed in low oxygen (5%) in a NAPCO 8000 WJ incubator (Thermo Scientific). Medium was equilibrated in 5% oxygen for approximately 4 h before use and cultures were passaged with TrypLE Select recombinant protease (Invitrogen) on days five and six. The daily RNA dose applied in the RiPSC derivations was 1,200 ng per well (six-well plate format). On day 21,

RiPS colonies were mechanically picked and transferred to MEF-coated 24-well plates with standard hESC medium (DMEM/F12 containing 20% KOSR (Invitrogen), 10 ng/ml bFGF (Gembio), 1x NEAA (Invitrogen), 0.1 mM b-ME (Sigma), 1 mM l-glutamine (Invitrogen), 50 units/ml penicillin and 50 µg/ml streptomycin) containing 5 mM Y27632 (BioMol). Clones were mechanically passaged once more to MEF-coated six-well plates, and then expanded via enzymatic passaging with collagenase IV (Invitrogen). Genomic DNA was extracted with a Qiagen DNeasy kit at the following passages: 9 (CF-Fib), 5 (CF-RiPS1.4), 5 (CF-RiPS1.9).

3.3.11 FiPS4F2 and FiPS4F-shpRb4.5 plasmid construction

pMX-Oct4, pMX-SOX2, pMX-KLF4, pMX-cMyc and pLVTHM were obtained from Addgene (plasmids 17217, 17218, 17219, 17220 and 12247, respectively). For the generation of the mammalian lentiviral plasmid encoding small hairpin RNAs against pRb-specific oligonucleotides were annealed, phosphorylated with T4 kinase and ligated into MluI/ClaI-linearized pLVTHM plasmid. The design of the small hairpin RNA was carried out using the SFOLD software (http://sfold.wadsworth.org/). All constructs generated were subjected to direct sequencing to rule out the presence of mutations.

3.3.12 FiPS4F2 and FiPS4F-shpRb4.5 retroviral and lentiviral production

Moloney-based retroviral vectors (pMX-) were co-transfected with packaging plasmids (pCMV-gag-pol-PA and pCMV-VSVg) in 293T cells using Lipofectamine (Invitrogen). Retroviral supernatants were collected 24 h after transfection, and passed through a 0.45 mM filter. Second-generation lentiviral vectors (pLVTHM-) were co-transfected with packaging plasmids (psPAX2 and pMD2.G, obtained from

Addgene, 12260 and 12259, respectively) in 293T cells using Lipofectamine (Invitrogen). Lentiviral supernatants were collected 36 h after transfection.

3.3.13 FiPS4F2P9, FiPS4F2P40 and FiPS4F-shpRb4.5 derivation

For the formation of hiPS cells, IMR90 fibroblasts were infected with equal proportions of retroviruses encoding for OCT4, SOX2, KLF4 and c-MYC plus empty lentiviruses (used to generate the FiPS4F2 line) or lentiviruses encoding small hairpin RNA against pRb (used to generate the line FiPS4F-shpRb4.5) by spinfection of the cells at 1,850 r.p.m. for 1 h at room temperature in the presence of polybrene (4 μg/ml). After two serial infections, cells were passaged onto fresh MEFs and switched to hES cell medium (DMEM/F12 (Invitrogen) supplemented with 20% Knockout serum replacement (Invitrogen), 1 mM l-glutamine, 0.1 mM non-essential amino acids, 55 mM β-mercaptoethanol and 10 ng/ml bFGF (Joint Protein Central)) four days after the first infection. For the derivation of hiPS cell lines, colonies were manually picked and maintained on fresh MEF feeder layers for five passages before the growth in Matrigel/mTesR1 (Stem Cell Technologies) conditions. DNA was extracted after nine passages for FiPS4F2P9 and FiPS4F-shpRB4.5 and 40 passages for FiPS4F2P40.

3.3.14 FiPS4F2 and FiPS4F-shpRb4.5 characterization

Cell pellets were lysed in 10 mM Tris-HCl (pH 8), 150 mM NaCl, 1% Triton X100, 1 mM Na3VO4, 1 mM PMSF and the Complete protease inhibitor mixture (Roche). Total protein extracts (25 μg) were used for SDS–PAGE, transferred to nitrocellulose membranes (Amersham Biosciences) and analyzed using primary antibodies against OCT4 (sc-5279, Santa Cruz), SOX2 (AB5603, Chemicom), RB1 (554136, Pharmingen) and Tubulin (T5168, Sigma). Horseradish-peroxidase-

conjugated secondary anti-mouse or rabbit were purchased from Cell Signaling and used at 1:5,000 dilution. Tubulin was used as a loading control. Immunoblots were visualized using SuperSignal solutions following the manufacturer's instructions (Thermo Scientific). Total RNA was isolated using TRIzol Reagent (Invitrogen), and cDNA was synthesized using the SuperScript II Reverse Transcriptase kit for RT–PCR (Invitrogen). Real-time PCR was performed using the SYBR-Green PCR Master mix (Applied Biosystems). Values of gene expression were normalized using GAPDH expression and are shown as fold change relative to the value of the sample control. All the samples were done in triplicate. The hiPS cell lines were cultured in the presence of 20 ng/ml colcemid for 45 min. The cells were trypsinized, washed with PBS and resuspended in a hypotonic solution by drop-wise addition while vortexing at low speed. After 10 min of incubation at 37 °C, cells were fixed by drop-wise addition of 1 ml of cold Carnoy's fixative. Stained metaphases were analyzed with CYTOVISION software (Applied Imaging). Teratoma analyses were performed as described[62].

3.3.15 Preparation of padlock probes

The design and preparation of padlock probes was based on published methods[10, 11, 63]. Libraries of long oligonucleotides (140 nucleotides) that cover different exonic regions were synthesized from programmable microarrays (Agilent Technologies). The libraries were amplified by performing 48–96 PCR reactions (100 µl each) with 0.02 nM template oligonucleotides, 200 nM Ap1V4IU primer (G*T*AGACTGGAAGAGCACTGTU), 200 nM Ap2V4 primer (/5Phos/TAGCCTCATGCGTATCCGAT), 0.2x SybrGreen I and 50 µl Econo Taq PLUS master mix (Lucigen), at 94 °C for 2 min; 17 cycles of 94 °C for 30 s, 58 °C for

30 s, 72 °C for 30 s; and finally 72 °C for 3 min. The amplicons were then purified by ethanol precipitation. Libraries were then digested with 40 units of Lambda Exonuclease (5 U/µl, NEB) in 1x Lambda Exonuclease buffer (NEB) at 37 °C for 2 h, followed by purification with four Qiagen Qiaquick PCR purification columns for every 48 wells of PCR products. Approximately 8 µg of the purified PCR amplicons were digested with ten units of DpnII (50 U/µl) in 1x DpnII buffer at 37 °C for 2 h, followed by the addition of four units of USER enzyme (1 U/µl, NEB) at 37 °C for 4 h. The DNA was digested with 6% PAGE and purified into single-stranded, 102-nucleotide probes.

3.3.16 Multiplex capture of exonic regions

Padlock probes (600 nM total concentration), 250 ng of genomic DNA, 1 nM suppressor oligonucleotides and 1x Ampligase buffer (Epicentre) were mixed in a 15-µl reaction and denatured at 95 °C for 10 min, then gradually cooled at the rate of 0.1 °C/s to 60 °C. The mixture was hybridized at 60 °C for 24 h. To circularize the captured targets, the reactions were then incubated at 60 °C for another 24 h after adding 2 µl of gap-filling mix (two units of AmpliTaq Stoffel (Life Technology), four units of Ampligase (Epicentre), and 500 pmol total dNTP). After circularization, 2 µl of exonuclease mix containing 10 U/µl exonuclease I (USB) and 100 U/µl exonuclease III (USB) was added to digest the linear DNA, and the reactions were incubated at 37 °C for 2 h and then inactivated at 94 °C for 5 min.

3.3.17 Amplification of capture circles

The 15-µl circularization products were placed in 100-µl PCR reactions with 200 nM of each primer (NH2-CAGATGTTATCGAGGTCCGAC, NH2-GGAACGATGAGCCTCCAAC, 0.2x SybrGreen I and 1x Phusion High-Fidelity PCR Master Mix (NEB) at 98 °C for 1 min; 16 cycles of 98 °C for 10 s, 58 °C for 20 s, 72 °C

for 20 s; and 72 °C for 3 min. The amplicons of the expected size range (200 bp) were purified using Qiagen Qiaquick columns.

3.3.18 Shotgun sequencing library construction

Purified PCR products with the four probe sets on the same template DNA were pooled in equal molar ratio. The PCR products were transferred into Covaris microTubes with snap caps for Covaris AFA shearing using a 10% duty cycle, an intensity setting of 5 and 200 cycles per burst. The sheared DNA was concentrated to 85 µl using a vacufuge, and was then prepared for sequencing library construction using NEBNext DNA Sample Prep Master Mix Set 1 (NEB). The fragmented DNA was end-repaired at room temperature for 30 min in 100-µl reaction consisting of 1x NEBNext End Repair Reaction Buffer and 5 µl of NEBNext End Repair Enzyme Mix. The DNA was then purified with Qiagen Qiaquick columns. Approximately 500 ng to 1 µg of the end-repaired blunt DNA was incubated in a thermal cycler for 30 min at 37 °C along with 1x NEBNext dA-Tailing Reaction Buffer and 3 µl of Klenow fragment. The DNA was again purified using Qiagen Qiaquick columns. The purified DNA was size-selected (125–150 nucleotides) using E-Gel SizeSelect 2% (Invitrogen) and concentrated to 36 µl using a vacufuge (Eppendorf). The dA-tailed DNA was then ligated at room temperature for 15 min with 1x Quick Ligation Reaction Buffer, 1.6 nM Illumina ligation adaptors and 2 µl of Quick T4 DNA ligase. Ligation products were purified using Qiagen Qiaquick columns and amplified by PCR in 100-µl reactions with a 15-µl template, 200 nM Illumina PCR primers, 0.2x SybrGreen I and 1x Phusion High-Fidelity PCR Master Mix (NEB) at 98 °C for 1 min; eight cycles at 98 °C for 10 s, 65 °C for 20 s, 72 °C for 15 s; and 72 °C for 3 min. The PCR amplicons were

purified with Qiaquick PCR purification columns, size-selected (200–275 nucleotides) using 6% PAGE and sequenced on an Illumina Genome Analyzer IIx.

3.3.19 Hybridization capture with DNA or RNA baits

Liquid exome capture was performed using the commercial Roche NimbleGen SeqCap EZ Exome kit or the commercial Agilent SureSelect kit (Table 3.1). Experiments were performed following the manufacturers' protocols. Briefly, genomic DNA was sheared and ligated to Illumina sequencing adaptors. DNA was then hybridized with the SeqCap EZ Exome library or SureSelect RNA baits to capture exomic regions. Exome regions were captured with streptavidin beads and then PCR-amplified with Illumina sequencing adaptors. The resulting libraries were sequenced on an Illumina Genome Analyzer IIx.

3.3.20 Consensus sequence generation and variant calling

Reads obtained from the Illumina Genome Analyzer were post-processed and quality filtered using GERALD. The end of each read was then mapped to the padlock-probe capturing arm sequences using Bowtie; any reads that successfully mapped were discarded to prevent bias from capturing arms. Reads were then mapped to the whole genome using Bowtie or BWA. Any read that could not be mapped uniquely was discarded to reduce false positives due to sequence homology. The 5′ and 3′ ends of reads were then trimmed to reduce the effect of sequencing errors, which tend to occur near the beginnings and ends of reads on the Illumina platform. (No trimming was performed when GATK was used for variant calling.) To reduce errors introduced by pre-sequencing amplification, mapped reads that started and ended at identical locations were then removed using SamTools or Picard to account for these clonal reads. SamTools or GATK was then used to generate a

consensus sequence for each sample by combining the results of each read that mapped to each exomic location. A minimum read depth of eight and consensus quality of 30 was required at every examined location. The consensus sequences were then compared to look for candidate novel mutations in hiPS cells. Variants that occurred at locations present in the dbSNP database (version 130) were removed from consideration to reduce the false-positive rate, as a novel mutation in the hiPS cell line is very unlikely to have been previously characterized in other cell lines and was most probably just not observed in the fibroblast line owing to stochastic sequencing bias. Because sequencing depth was relatively low in a small fraction of exomic regions, allelic imbalance can also lead to false positives, as sites in the fibroblast genome could, for example, be heterozygous but be sequenced as seven copies of the major allele and one copy of the minor allele and called as homozygous. To prevent these false positives, sites in which the fibroblast genome showed even a very small presence of minor allele were removed from consideration as candidate sites for novel mutations (as these sites are most probably truly heterozygous in both lines). Several locations were identified in which the hiPS cell sample consensus sequence showed a heterozygous call but the fibroblast sample consensus sequence showed a homozygous call; these were identified as candidate mutations, as it is expected that during mutational processes, the hiPS cell sample would most probably gain an additional allele. These candidate mutations were then validated by capillary sequencing as below.

3.3.21 Sanger validation of candidate mutations

Genomic DNA (6 ng) was amplified in a 50-µl PCR reaction with 100 nM specifically designed primers near the mutation site and 25 µl Taq 2x master mix (NEB) at 94 °C for 2 min; 35 cycles at 94 °C for 30 s, 57 °C for 30 s and 72 °C for 30 s;

and final extension at 72 °C for 3 min. The PCR products were then purified with Qiagen Qiaquick columns, and 10 ng of purified DNA was pre-mixed with 8 pmol of the forward sequencing primer for capillary Sanger sequencing by Genewiz.

3.3.22 Clonal fibroblast experiments

In an attempt to determine the mutational load present in single fibroblasts, we performed a reprogramming-like clonal colony purification strategy on fibroblasts. CV fibroblasts were thawed at passage 14 and cultured in fibroblast media (DMEM containing 15% FBS, penicillin/streptomycin, sodium pyruvate, non-essential amino acids and l-glutamine). A confluent 6-cm plate was trypsinized and cells were plated in three 96-well dishes, in the presence (two plates) or absence (one plate) of MEF feeder cells, at limiting dilutions. Another 96-well plate was plated as a reference plate. Using Poisson calculations, cells were diluted and plated such that it was extremely unlikely (<1%) for one well to contain more than one cell (leading to an expectation of eight wells per plate with one cell). These wells were cultured and progressively passaged from the 96-well dish to a 6-cm plate (96-well, 48-well, 24-well, 12-well, 6-well, 6-cm). For cells growing on MEFs, all passages from a 12-well dish to a 6-cm dish were done without MEFs to minimize contamination with mouse cells in the sequencing analysis. Only three MEF-free wells and nine MEF-containing wells successfully grew; using Poisson calculations, 24 wells should have successfully grown.

All fibroblasts grown from single cells showed heavy signs of stress. Cells grew very slowly (with passaging needed approximately every one to two weeks). MEF-free cells had a flattened morphology, whereas MEF-plated cells maintained a normal, spindle-shaped morphology. Cells tended to senesce very soon after plating; only a few cells grew successfully. Seven clonal lines were sequenced (three grown

without MEFs and four grown with MEFs). Six of the lines contained a very high number of putative mutation candidates (~100), and no mutations were found in one line grown on MEFs. We randomly selected 21 of the 600 mutation candidates for Sanger validation, and found that approximately 50% were true positives. This leads to a projection of ~50 protein-coding mutations in six clonal fibroblast lines, which is tenfold more than what was observed in hiPS cells and not consistent with the observations on the other clonal fibroblast line, which was completely mutation free. We proposed that the mutations in the six clonal fibroblast lines were due to the stress associated with expanding single fibroblast cells. Because fibroblast growing conditions are very different from those found in reprogramming, we cannot estimate the background somatic mutation rate in such an experiment. We therefore instead used published estimates of fibroblast mutation rate to estimate clonal fibroblast mutational load (see below).

3.3.23 Digital quantification of mutations

Thirty-two pairs of DigiQ-PCR primers were designed such that the forward or reverse primers are roughly 25 base pairs away from the 5′ end of each mutation site. This ensured that the mutations of interest were sequenced in the part of the read length that had the highest accuracy. Primers also contained an annealing region for Illumina Solexa sequencing primers at the 5′ ends. Each primer corresponding to a different mutation was amplified with a high-fidelity polymerase in three samples: the mutated hiPS cell line, the progenitor fibroblast line and a clean control. To sample DNA from 100,000 cells, 600 ng of DNA was used for each mutated hiPS cell line and fibroblast line. In cases where a separate clonal hiPS cell line not containing the mutation in question was available, this line was used as a clean control, as the chance of this line acquiring the same mutation during clonal expansion is extremely

low (~10−9 for one mutation). In other cases, a ‘low-input’ sample using 300 pg of DNA (~50 cells) was used, as rare mutations are unlikely to be present in such a small quantity of DNA. If any mutated DNA was sampled, it would be immediately obvious in the sequencing results and the experiment could be repeated. First-round PCR amplification was performed with 600 ng (~100,000 cells) of DNA, 500 nM of each DigiQ-PCR primer and 1x iProof High-Fidelity Master Mix (Bio-Rad) at 98 °C for 30 s; ten cycles at 98 °C for 10 s, 59 °C for 20 s, and 72 °C for 15 s; 18 cycles at 98 °C for 10 s and 72 °C for 20 s; and final extension at 72 °C for 3 min. The PCR amplicons were purified using Qiaquick columns (Qiagen). Roughly 100 ng of the first-round PCR product was used as a template for second-round PCR amplification, together with 1x Phusion High-Fidelity PCR Master Mix (NEB) and 200 nM of each Illumina PCR primer, at 98 °C for 30s; ten cycles at 98 °C for 10 s and 64 °C for 30 s; and final extension at 72 °C for 30 s. The amplicons were purified again with Qiaquick columns (Qiagen) and size-selected (roughly 150–200 nucleotides) using an E-Gel SizeSelect 2% system (Invitrogen). PCR reactions were performed with the iProof High-Fidelity Master Mix (Bio-Rad) and Phusion High-Fidelity PCR Master Mix (NEB) to minimize amplification errors. All size-selected products were pooled together at equal ratio; these libraries were then mixed with the Illumina PhiX control library in a roughly equal ratio to balance the fluorescent signals at all four bases and improve the base-calling accuracy, and sequenced using an Illumina GA IIx. Each pair of libraries from the fibroblasts and negative controls was sequenced in two non-adjacent lanes of a same flow cell. Extreme care was taken in sample handling to ensure no cross-contamination from the positive control libraries to the other libraries. Alleles identified at each mutation position by the sequencer were counted and evaluated. The specific sample choices for each mutation are not shown to conserve space (for details, see

Supplementary Fig. 3.2 and original manuscript[36]). To verify the robustness of the DigiQ assay, the assay was repeated on CV fibroblasts. The obtained read proportions were extremely similar (Supplementary Fig. 3.3).

3.3.24 Statistical analysis—probability of mutations occurring naturally

We evaluated the likelihood that the mutations found were generated during fibroblast culturing and reprogramming (assuming a clean starting population of fibroblasts) at the normal estimated somatic mutation rate of between 10^{-6} and 10^{-7} non-synonymous coding mutations per gene per cell division, which corresponds to a rate of 6.7×10^{-10} (using the average human coding-region size of 1,500 base pairs per gene[64]). Assuming that mutations are independent events that occur uniformly across the genome, the number of expected mutations during fibroblast culturing and reprogramming can be estimated using a Poisson distribution with expected value $\lambda = 6.7 \times 10^{-10}ns$, where n is the number of cell divisions and s is the observed sequence. Although accurate records of the number of cell divisions experienced by each line during expansion and reprogramming are not available, we estimated that 30–35 doublings had occurred for six lines with well-documented culture histories. In these lines, a total of 206,227,380 base pairs were pairwise-sequenced (at a depth of at least eight and quality of at least 30). This leads to a Poisson distribution with $\lambda = 4.13–4.81$ for the expected number of mutations. In this case, we observed 54 coding mutations, leading to a P value of $1.29 \times 10^{-40}–2.72 \times 10^{-37}$. If this calculation is extrapolated to all 22 lines, we expect $\lambda = 8.7–10.1$ coding mutations; we observed 91, leading to a P value of $4.29 \times 10^{-59}–1.27 \times 10^{-53}$. We can therefore say that these mutations did not occur by chance with more than 99% confidence for all 22 lines.

3.3.25 Statistical analysis—digital quantification

To quantify the frequency of each mutation in the fibroblast samples, a one-tailed binomial distribution test was used. Reads were quality-filtered; only base calls with a Phred-like quality score of 30 or greater were considered. We denote by p the probability of obtaining a sequencing read containing the minor allele. The fibroblast sample was compared with either the clean low-input sample or a clean clonal hiPS cell line. Because the two hiPS cell lines are clonally independent, they will not share any mutations. Therefore, for example, FS-low can be used as a negative control for FS and CV-hiPS-B can be used as a negative control for CV-hiPS-F. Any minor allele obtained from the clonal hiPS cell or low-input fibroblast sample will be purely due to sequencing error. We denote by H0 the event that the minor allele frequency in the fibroblast sample was less than or equal to the minor allele frequency in the other clonal/low-input sample, and denote by H1 the event that the minor allele frequency in the fibroblast sample was greater. If H0 is found to be true, the mutation cannot be detected in the fibroblast, as any presence of the minor allele cannot be distinguished from sequencing error. If H1 is found to be true, the presence of the minor allele is detectable and can be quantified. We denote by n the total number of reads that called the mutated position. A critical value of a = 0.01 was chosen (99% confidence). Because the number of reads for each sample was very high, both np and n(1 − p) were greater than five, meaning that the minor allele presence could be approximated with a normal distribution. We can therefore set a criterion for rejection of the null hypothesis of $Z = (x - \mu)/s > 2.33$, where x is the minor allele count, μ is the mean of the minor allele counts of the fibroblast and low-input/clonal samples, and s is the standard deviation of the minor allele counts of the fibroblast and low-input/clonal samples. For a binomial-distribution approximation, n is the number of reads in the

fibroblast sample, p is the minor allele frequency if the fibroblast and low-input/clonal data are merged, μ = np, and s = np(1 − p). If the value of Z is greater than 2.33, we are capable of distinguishing the observed fraction of minor alleles in the fibroblast sample from that observed in the clonal/low-input sample. These results are presented in Supplementary Table 3.1.

We can also construct a 99% confidence interval using the normal approximation for the binomial distribution. Although we observed a value for the minor allele in each fibroblast sample, due to sequencing error, this value may overestimate or underestimate the true minor allele frequency. We can counteract this error using a normal distribution. The confidence-interval values are derived from the normal probability density function and represent the boundaries that we are 99% sure the true minor allele frequency lies within: lower bound, min((−2.57s + x)/n, 0); upper bound, min((2.57s + x)/n, 0). These estimates for the minor allele fraction in fibroblasts are shown in Supplementary Table 3.1.

3.3.26 Statistical analysis—NS/S mutation ratio

To determine whether selection pressure could have a role in reprogramming-associated mutations, we compared the mutational load associated with reprogramming with that associated with tumorigenesis. The NS/S ratio found in several previously conducted pairwise cancer sequencing analyses[65-67] was found to be 2.4:1. The load found here out of 124 identified mutations is 2.6:1, meaning that hiPS cell lines carry a very similar mutational pattern to cancer lines.

3.3.27 Statistical analysis—pathway and COSMIC gene enrichment

To check for enrichment of reprogramming-associated mutated genes in cancer-related genes, the fraction of genes mutated in hiPS cells found mutated in

the COSMIC[68] database was identified as 50/124. As 4,471 of the 16,017 genes well targeted by our exome sequencing pipeline are considered to be commonly mutated in cancer, a $\chi 2$ test with one degree of freedom can be used to test for equivalency of distribution. The obtained $\chi 2$ value is 9.67, indicating that the fraction of mutated hiPS cell genes in the COSMIC set is statistically significantly greater than the normally obtained number with a P value of 0.001873. This indicates that hiPS cell mutations are enriched in COSMIC set genes at approximately 1.5-fold the normal level, of 28%, with >99% confidence. To check for commonly mutated pathways, reprogramming-associated mutated genes and mutated genes identified in three cancer sequencing papers[65-67] were analyzed using DAVID[69]. No statistically significant pathway Gene Ontology terms were identified; the lowest Benjamini P value found was 0.6, which is well above the cut-off value, of 0.01, required for 99% confidence. Therefore, no common pathways seem to be mutated in hiPS cells.

3.4 Results

3.4.1 hiPS cell lines contain a high level of mutational load

We identified sites that showed the gain of a new allele in each hiPS cell line relative to their corresponding matched progenitor fibroblast genome. A total of 124 mutations were validated with capillary sequencing (Fig. 3.1, Table 3.2), which revealed that each mutation was fixed in heterozygous condition in the hiPS cell lines. No small insertions/deletions were detected. For three hiPS cell lines (CV-hiPS-B, CV-hiPS-F and PGP1-iPS), the donor's complete genome sequence obtained from whole blood is publicly available[70, 71]; we used this information to further confirm that all 27 mutations in these lines were bona fide somatic mutations. Because 84% of the expected exomic variants[72] were captured at high depth and quality, the predicted

load is approximately six coding mutations per hiPS cell genome (see Table 3.1 for details). The majority of mutations were mis-sense (83 of 124), nonsense (5 of 124) or splice variants (4 of 124). Fifty-three mis-sense mutations were predicted to alter protein function[73]. Fifty mutated genes were previously found to be mutated in some cancers[68, 74]. For example, ATM is a well-characterized tumor suppressor gene found mutated in one hiPS cell line, and NTRK1 and NTRK3 (tyrosine kinase receptors) can cause cancers when mutated[75] and contained damaging mutations in three hiPS cell lines (CV-hiPS-F, iPS29e and FiPS4F-shpRB4.5) that were reprogrammed in three labs and came from different donors. Two kinase genes from the NEK family, which is related to cell division, were mutated in two independent hiPS cell lines. In addition to cancer-related genes, 14 of the 22 lines contained mutations in genes with known roles in human Mendelian disorders[76]. Three pairs of hiPS cell lines (iPS17a and iPS17b, dH1F-iPS8 and dH1F-iPS9, and CF-RiPS1.4 and CF-RiPS1.9) shared three, two and one mutation, respectively; these most probably arose in shared common progenitor cells before reprogramming. However, most hiPS cell lines derived from the same fibroblast line did not share common mutations (Table 3.2).

These data raise the possibility that a significant number of mutations occur during or shortly after reprogramming and then become fixed during colony picking and expansion. An alternative hypothesis is that the mutations we found are simply the result of age-accrued biopsy heterogeneity or normal somatic mutation during in vitro fibroblast cell culture. The skin biopsies were collected from donors of ages varying from newborn to 82 years; biopsy heterogeneity therefore does not seem to have a primary role, as the mutational load is not correlated (squared linear correlation coefficient, $R^2 = 0.046$) with donor age (Supplementary Fig. 3.1). We attempted to grow clonal fibroblasts to obtain a control for single-cell mutational load,

but a direct assessment was not possible owing to technical difficulties in mimicking the exact culture conditions (Methods). Assuming that the skin biopsy is mutation free, we were able to use previously published values for the typical mutation rate in culture to obtain an expectation of ten times fewer mutations per genome than we observed ($P < 1.27 \times 10^{-53}$; Methods), indicating that hiPS cell mutational load is higher than normal-culture mutational load. We define the term 'reprogramming-associated mutations' to describe somatic mutations observed in these hiPS cell lines. Reprogramming-associated mutations could pre-exist at low frequencies in the fibroblast population, could occur during the reprogramming process or could occur after reprogramming. All reprogramming-associated mutations have become fixed in the hiPS cell population.

3.4.2 Reprogramming-associated mutations arise through multiple mechanisms

To test whether some observed mutations were present in the starting fibroblasts at low frequency before reprogramming, we developed a new digital quantification assay (DigiQ) to quantify the frequencies of 32 mutations in six fibroblast lines using ultra deep sequencing (Supplementary Figs. 3.3 and 3.4). We amplified each mutated region from the genomic DNA of 100,000 cells with a high-fidelity DNA polymerase and sequenced the pooled amplicons with an Illumina Genome Analyzer at an average coverage of 10^6. Although the raw sequencing error is roughly 0.1–1% with the Illumina sequencing platform, detection of rare mutations at a lower frequency is possible with proper filtering and careful selection of controls[77]. For each fibroblast line, we included the mutation-carrying hiPS cell DNA as the positive control and a 'mutation-free' DNA sample as the negative control for sequencing errors (Methods). Comparison of the allelic counts at the mutation

positions between the fibroblast lines and the negative controls allowed us to distinguish rare mutations from sequencing errors and estimate the detection limit of the assay. Seventeen of the 32 mutations were found in fibroblasts in the range of 0.3–1,000 in 10,000, and 15 mutations were not detectable (Supplementary Tables 3.2 and 3.3). In each fibroblast line with more than one detectable rare mutation, the frequencies of the mutations were very similar, which suggests that a small subpopulation of each fibroblast line contains all pre-existing hiPS cell mutations and that the rest of the cells lacked any of them.

We extended this analysis by asking whether all of the hiPS cell mutations could have pre-existed in the fibroblast populations. For the 15 mutations not detected with the DigiQ assay, the detection limits can be estimated (Methods). At seven of the 15 sites, the sequencing quality was high enough that rare mutations at frequencies of 0.6–5 in 100,000 should be detectable with our assay (Supplementary Table 3.1). Because 30,000–100,000 fibroblast cells were used in the reprogramming experiments, we can rule out the presence of two mutated genes (NTRK3 and POLR1C) in more than one cell of the starting fibroblast population, and five others were present in no more than one or two cells.

As another test of the hypothesis that all of the mutations pre-existed in fibroblasts before reprogramming, we examined the exomes of two hiPS cell lines derived from fibroblast line dH1cf16, which was clonally derived from the dH1F fibroblast line and passaged the minimum amount to generate enough cells for reprogramming. The two hiPS cell lines derived from the non-clonal dH1F fibroblast line contained eight and three, respectively, new mutations not found in the fibroblasts; we observed a very similar independent mutational load in the clonal lines (six new mutations in the hiPS cell line dH1cf16-iPS1 and two new mutations in the

hiPS cell line dH1cf16-iPS4). Together, these experiments establish that although some of the reprogramming-associated mutations were likely to pre-exist in the starting fibroblast cultures, the others occurred during reprogramming and subsequent culturing. Specific distributions tend to vary across hiPS cell lines (Supplementary Table 3.1).

Mutations that occur during reprogramming could be due in part to a significantly elevated mutation rate during reprogramming. It is also possible that selection could have an important role. We tested the possibility that an elevated mutation rate might occur because the reprogramming process might be inducing transient repression of p53 (also known as TP53), RB1 and other tumor suppressor genes, which are known to inhibit reprogramming and are required for normal DNA damage responses. Simian virus 40 large-T antigen, which inactivates tumor suppressor and DNA damage response genes[78] (including p53 and RB1), was expressed during reprogramming of three analyzed hiPS cell lines (DF6-9-9, DF19-11 and iPS4.7)[79]. Another hiPS cell line (FiPS4F-shpRB4.5) was generated while directly knocking down RB1 (Supplementary Fig. 3.4). However, the observed mutational load was very similar in these lines in comparison with the others, indicating that reprogramming-associated mutations cannot be explained by an elevated mutation rate caused by p53 or RB1 repression.

We also probed whether additional mutations could become fixed during extended passaging by extending our analysis of one hiPS cell line. Although most of our hiPS cell lines were sequenced at fairly low passage number (less than 20), to measure the effect of post-reprogramming culturing directly we also sequenced one hiPS cell line (FiPS4F2) at two passages (9 and 40). We discovered that all seven

mutations identified in the passage-9 line remained fixed in the passage-40 line, but that four additional mutations were found to be fixed in the passage-40 cell line.

To test the possibility that selection operates during hiPS cell generation, we performed an enrichment analysis to determine whether reprogramming-associated mutated genes were more likely to be observed than random somatic mutation in cancer cells. We used the COSMIC database as a source of genes commonly mutated in cancer[68]. We discovered that the reprogramming-associated mutated genes were significantly enriched for genes found mutated in cancer ($P = 0.0019$; Supplementary Information), which implies that some mutations were selected during reprogramming.

As an alternative test of the selection hypothesis, we asked whether mutations associated with reprogramming could be functional, on the basis of the non-synonymous/synonymous (NS/S) ratio. Traditionally, the analysis of the NS/S ratio is applied to germline mutations that have evolved over a long period of evolutionary time, and is not directly applicable to somatic mutations. However, functional mutations are known to be positively selected in cancers, allowing us to make a direct comparison with mutation characteristics found in cancer genomes. Strikingly, the NS/S ratio is very similar between mutations identified in three recent cancer genome sequencing projects[65-67] and the reprogramming-associated mutations we found (2.4:1 and 2.6:1, respectively), indicating that a similar degree of selection pressure may be present.

We also checked whether reprogramming-associated mutations could provide a common functional advantage, through a pathway enrichment analysis using Gene Ontology terms[69]. No statistically significant similarity was identified, indicating that mutated genes have varied cellular functions. Again, identical results were found

when performing the same analysis on mutations identified during the genome sequencing of melanoma, breast cancer and lung cancer samples[65-67]. This lack of enrichment in cancer genomes is generally thought to be due to the presence of many passenger mutations in cancer cells, which could also be the cause for reprogramming-associated mutations. Nonetheless, these analyses suggest that selection of potentially functional mutations could have a role in amplifying rare-mutation-carrying cells and, when coupled with the single-cell bottleneck in hiPS cell colony picking, could contribute to the fixation of initially low-frequency mutations throughout the entire hiPS cell population.

3.5 Conclusions

Taken together, our results demonstrate that pre-existing and new mutations that occur during and after reprogramming all contribute to the high mutational load we discovered in hiPS cell lines. Although we cannot completely rule out the possibility that reprogramming itself is 'mutagenic', our data argue that selection during hiPS cell reprogramming, colony picking and subsequent culturing may be contributing factors. A corollary is that if reprogramming efficiency is improved to a level such that no colony picking and clonal expansion is necessary, the resulting hiPS cells could potentially be free of mutations.

Despite the power of our experimental approach to identify and characterize reprogramming-associated mutations accurately, their functional significance remains to be shown. This issue parallels a general problem facing the genomics community: high-throughput sequencing technologies have allowed data generation rates to greatly outpace functional interpretation. Additionally, when considering the biological significance of reprogramming-associated mutations, there are two separate

functional aspects to consider: whether some of these mutations contributed functionally to the reprogramming of cell fate, and whether some of these mutations could increase disease risk when hiPS-cell-derived cells/tissues are used in the clinic. These two aspects are not necessarily connected. Although the functional effects of the 124 mutations remain to be characterized experimentally, it is nonetheless striking that the observed reprogramming-associated mutational load shares many similarities with that observed in cancer. Furthermore, the observation of mutated genes involved in human Mendelian disorders suggests that the risk of diseases other than cancer needs to be evaluated for hiPS-cell-based therapeutic methods. Future long-term studies must focus on functional characterization of reprogramming-associated mutations to aid further the creation of clinical safety standards.

Safe hiPS cells are critical for clinical application. Therefore, just as previous findings of large-scale genome rearrangements in hiPS cell lines led to the introduction of karyotyping as a standard post-reprogramming protocol, routine genetic screening of hiPS cell lines to ensure that no obviously deleterious point mutations are present must become a standard procedure. Complete exome or genome sequencing of hiPS cell lines might be an efficient way to screen out hiPS cell lines that have a high mutational load or have mutations in genes implicated in development, disease or tumorigenesis. Further rigorous work on mutation rates and distributions during in vitro culturing and reprogramming of hiPS cells, and perhaps human embryonic stem cells, will be essential to help establish clinical safety standards for genomic integrity.

3.6 Acknowledgements

We thank J. M. Akey, G. M. Church, S. Ding, J. B. Li and J. Shendure for discussions and suggestions, S. Vassallo for assistance with DNA shearing, and G. L. Boulting and S. Ratansirintrawoot for assistance with hiPS cell culture. This study is supported by NIH R01 HL094963 and a UCSD new faculty start-up fund (to K.Z.), a training grant from the California Institute for Regenerative Medicine (TG2-01154) and a CIRM grant (RC1-00116) (to L.S.B.G.). L.S.B.G. is an Investigator of the Howard Hughes Medical Institute. A. Gore is supported by the Focht-Powell Fellowship and a CIRM predoctoral fellowship. M.L.W. is supported by an institutional training grant from the National Institute of General Medical Sciences (T32 GM008666). Y.-H.L. is supported by the A*Star Institute of Medical Biology and the Singapore Stem Cell Consortium. Work in the laboratory of J.C.I.B. was supported by grants from MICINN, Sanofi-Aventis, the G. Harold and Leila Y. Mathers Foundation and the Cellex Foundation. G.Q.D. is an investigator of the Howard Hughes Medical Institute and supported by grants from the NIH.

Chapter 3, in part, is a reprint of the material as it appears in: Athurva Gore*, Zhe Li*, Ho-Lim Fung, Jessica E. Young, Suneet Agarwal, Jessica Antosiewicz-Bourget, Isabel Canto, Alessandra Giorgetti, Mason A. Israel, Evangelos Kiskinis, Je-Hyuk Lee, Yuin-Han Loh, Philip D. Manos, Nuria Montserrat, Athanasia D. Panopoulos, Sergio Ruiz, Melissa L. Wilbert, Junying Yu, Ewen F. Kirkness, Juan Carlos Izpisua Belmonte, Derrick J. Rossi, James A. Thomson, Kevin Eggan, George Q. Daley, Lawrence S. B. Goldstein, Kun Zhang. “Somatic coding mutations in human induced pluripotent stem cells.” *Nature*. 2011 March 3; 471: 63-67. doi:10.1038/nature09805. Used with permission. The dissertation author was one of the primary investigators and authors of this paper.

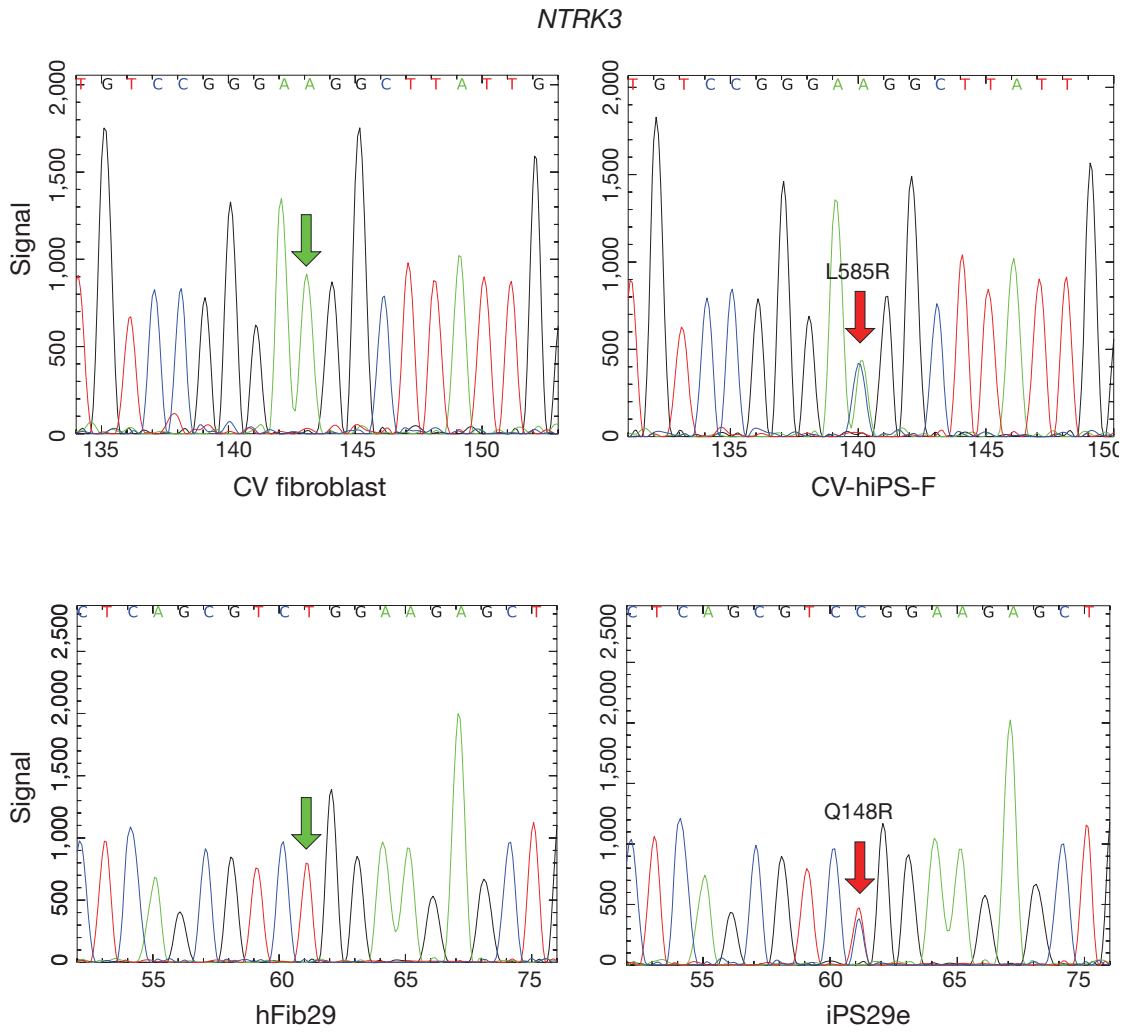
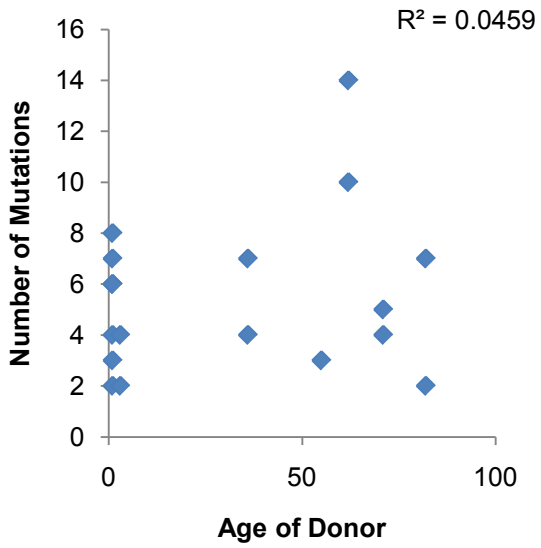
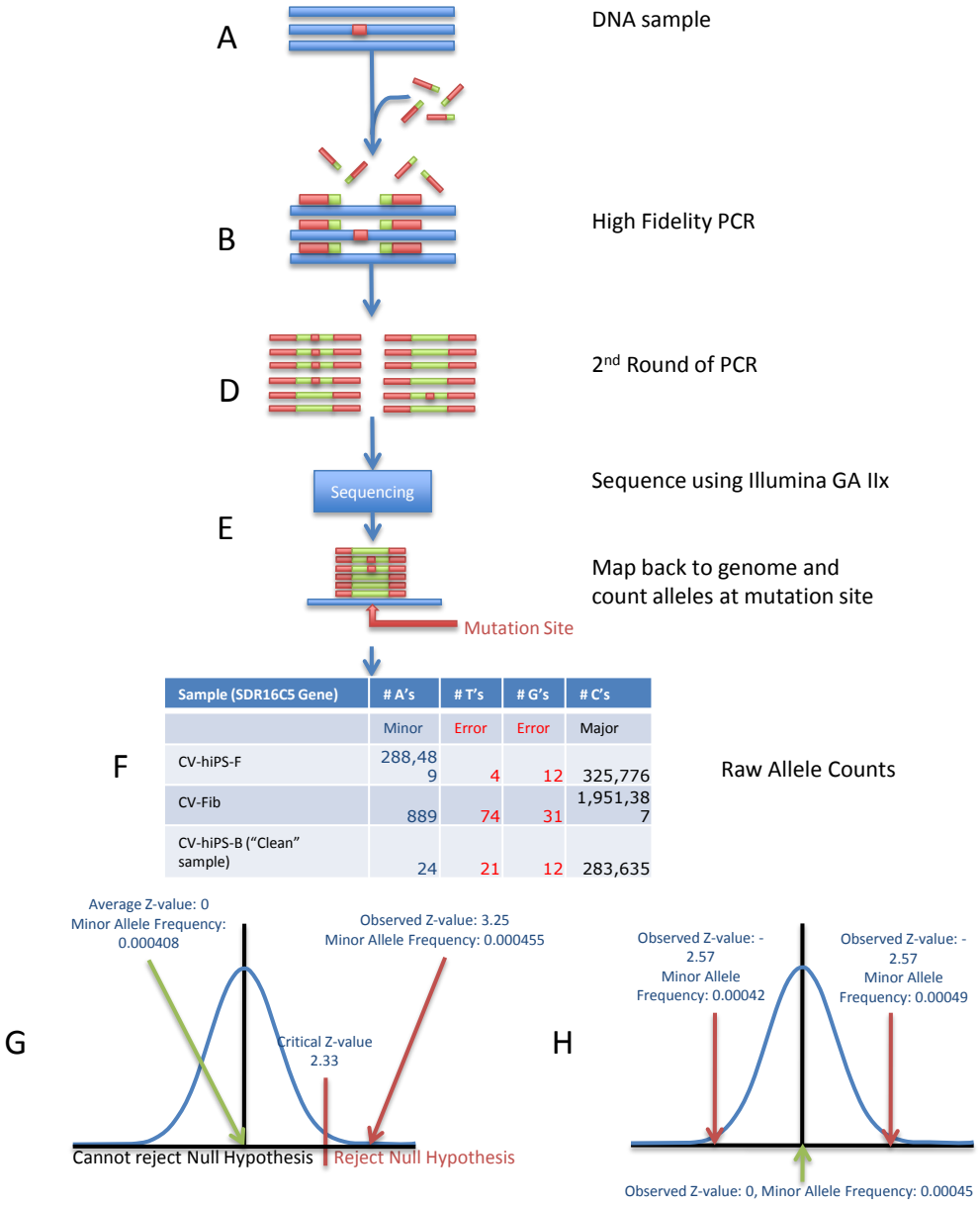


Figure 3.1. hiPS cells acquired protein-coding somatic mutations. Somatic mutations in the gene NTRK3 were found in two independent hiPS cell lines but were not present in their fibroblast progenitors.

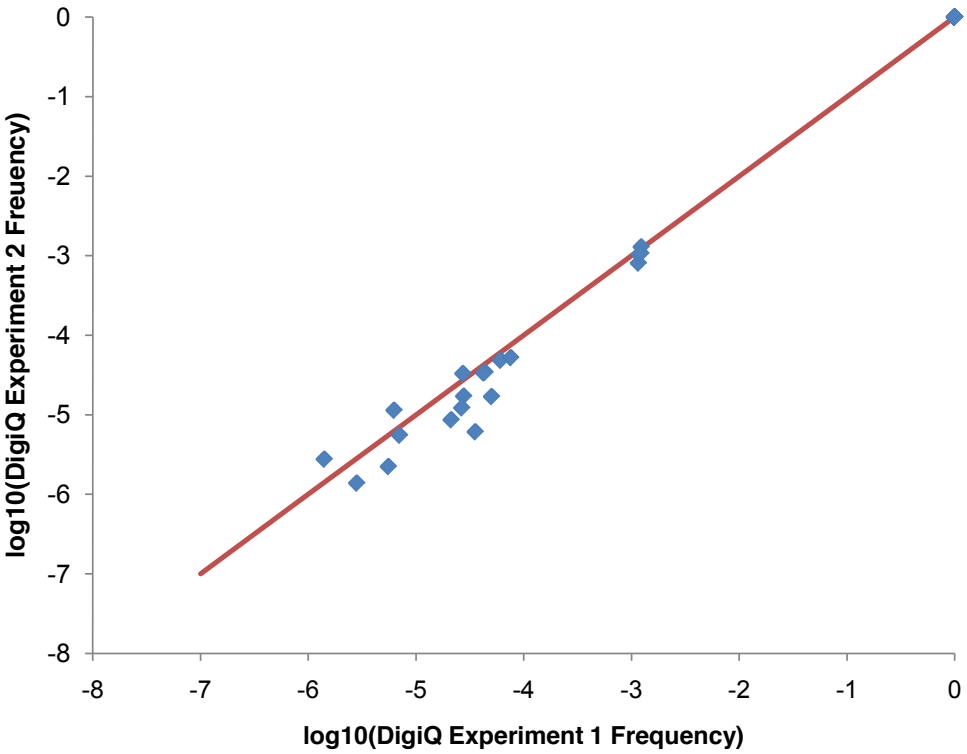


Supplementary Figure 3.1. Donor age versus mutation count. Mutational load was compared across the eighteen hiPS lines for which donor age is known. No correlation (R^2 = 0.046) was observed between donor age and mutational load, indicating that age-accrued mutation does not play a primary role in reprogramming-associated mutations.

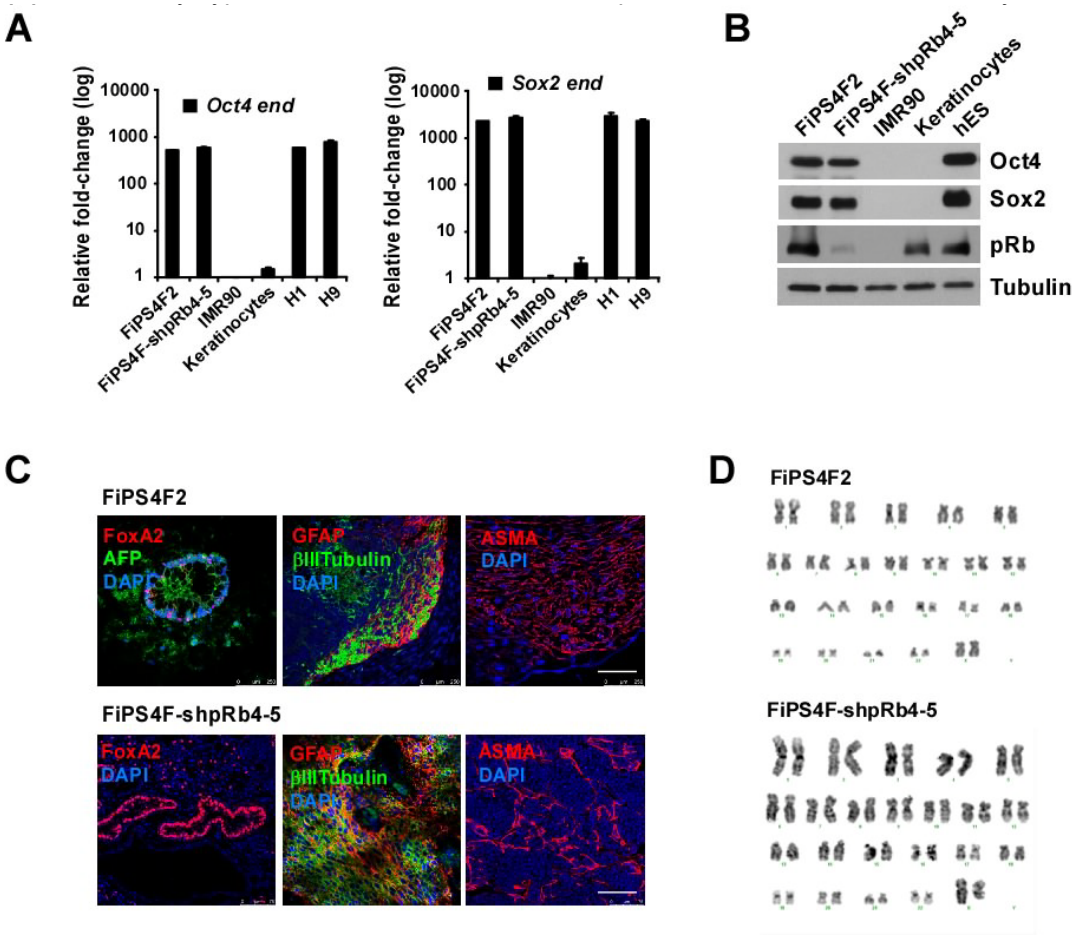


Sample (SDR16C5 Gene)	# A's	# T's	# G's	# C's
	Minor	Error	Error	Major
CV-hiPS-F	288,489	4	12	325,776
CV-Fib	889	74	31	1,951,387
CV-hiPS-B ("Clean" sample)	24	21	12	283,635

Supplementary Figure 3.2. Digital Quantification Experiment. (A) The initial DNA sample from a group of fibroblast cells contains some normal and some mutated genes. (B-C) PCR with complementary sequences (green) attached to Illumina sequencing tails (red) can be used to amplify the mutated region. (D) Further amplification with sequencing adaptors results in an Illumina GA IIx sequencing library that is then sequenced. (E) The reads resulting from this sequencing can be mapped back to the gene; the mutated position is indicated. (F) The fraction of each allele found at the mutated location can then be quantified. However, sequencing errors cause all four alleles to be present in some small quantity. (G) A normal approximation to the binomial distribution was used to check if the observed minor allele count could be distinguished from error. (H) A 99% confidence interval was constructed to find the upper and lower bounds of mutation frequency.



Supplementary Figure 3.3. Robustness of Digital Quantification (DigiQ) Assay. DigiQ was performed twice in CV-Fibroblasts. The log (base 10) of allele counts from experiment 1 are plotted below versus those from experiment 2. The red line is y=x, corresponding to identical results. The DigiQ assay produces very consistent results from experiment to experiment.



Supplementary Figure 3.4. Characterization of Rb knockout iPSC line. (A) Real-time PCR analysis of the relative expression of the Oct4 and Sox2 endogenous levels in the hiPS cell lines (FiPS4F2 and FiPS4F-shpRb4-5) compared to hES cells (H1 and H9 cell lines), keratinocytes and the somatic cell of origin (IMR90). Data are shown as the relative averages ± SEM calculated from two biological replicates analyzed in triplicate. (B) Western blot analysis of the noted proteins in the same cell lines described in (A). Note the efficient downregulation in the levels of pRb in the FiPS4F-shpRb4-5 cell line. (C) Teratoma formation was assessed by injection of the FiPS4F2 and FiPS4F-shpRb4-5 cell lines into the testes or kidney of SCID mice. Immunofluorescence analysis demonstrate the existence of the three main embryonic germ layers as defined by the expression of specific endodermal (AFP (α-fetoprotein), and FoxA2), ectodermal (βIIITubulin and GFAP) and mesodermal (ASMA (alpha-smooth muscle actin)) markers. All images were obtained from the same tumor. Scale bar: 250μM. (D) Normal karyotype of the FiPS4F2 and FiPS4F-shpRb4-5 cell lines demonstrated by G-banding karyotype analysis.

Table 3.1. Sequencing statistics for mutation discovery.

Cell line	Exome capture method	Quality-filtered sequence (bp)	No. of high-quality coding variants	dbSNP percentage	Shared high-quality coding region (bp)	No. of coding mutations observed/projected
CV-hiPS-F	Padlock + SeqCap EZ	9,928,014,640	15,595	98%	16,374,878	14/15
CV-hiPS-B	SeqCap EZ	7,977,894,480	14,876	98%	21,891,518	10/12
CV fibroblast	Padlock + SeqCap EZ	7,586,731,600	15,442	98%	—	—
DF-6-9-9	Padlock + SeqCap EZ*	9,289,593,520	14,366	95%	17,806,151	6/7
DF-19-11	SeqCap EZ	3,212,662,880	13,792	95%	21,342,017	7/9
iPS4.7	SeqCap EZ	3,132,462,400	14,154	95%	21,729,562	4/5
Foreskin fibroblast	Padlock + SeqCap EZ*	8,430,654,720	14,819	95%	—	—
PGP1-iPS	SeqCap EZ	4,599,556,400	14,105	95%	19,681,915	3/4
PGP1 fibroblast	SureSelect	3,504,437,120	14,781	95%	—	—
dH1F-iPS8	SeqCap EZ	3,950,994,160	13,552	96%	16,874,057	8/10
dH1F-iPS9	SeqCap EZ	3,945,196,800	14,191	95%	21,536,158	3/4
dH1F fibroblast	SeqCap EZ	3,373,535,920	13,838	95%	—	—
iPS11a	SureSelect	1,836,303,440	13,845	95%	18,557,098	4/5
iPS11b	SureSelect	3,378,603,200	15,152	95%	17,206,934	7/8
Hib11 fibroblast	SureSelect	5,660,864,960	13,579	95%	—	—
iPS17a	SureSelect	4,805,756,800	15,039	95%	17,888,773	4/5
iPS17b	SureSelect	7,129,037,520	15,400	95%	19,902,076	5/6
Hib17 fibroblast	SureSelect	3,962,506,880	13,365	96%	—	—
iPS29A	SureSelect	4,112,237,360	13,464	94%	17,328,182	2/3
iPS29e	SureSelect	1,669,916,080	13,800	94%	18,985,791	7/9
Hib29 fibroblast	SureSelect	4,388,388,320	14,445	95%	—	—
dH1cF16-iPS1	SeqCap EZ	4,321,661,440	15,061	95%	19,601,528	2/2
dH1cF16-iPS4	SeqCap EZ	4,668,085,920	14,958	95%	23,956,732	6/7
dH1cF16 fibroblast	SeqCap EZ	4,178,664,160	14,879	95%	—	—
CF-RiPS1.4	SeqCap EZ	4,733,743,840	11,344	96%	21,272,233	2/3
CF-RiPS1.9	SeqCap EZ	3,143,591,760	13,674	95%	21,165,013	5/6
CF fibroblast	SeqCap EZ	3,204,874,880	11,855	96%	—	—
FiPS3F1	SeqCap EZ	3,397,397,360	13,333	94%	20,723,620	4/5
FiPS4F7	SeqCap EZ	3,346,801,280	14,584	94%	21,608,258	2/3
HFFXF fibroblast	SeqCap EZ	3,331,494,880	13,040	94%	—	—
FiPS4F2p9	SeqCap EZ	4,725,258,400	18,033	92%	25,188,054	7/7
FiPS4F2p40	SeqCap EZ	4,848,006,000	18,376	92%	25,411,595	11/11
FiPS4F-shpRB4.5	SeqCap EZ	4,911,008,400	19,491	92%	25,240,944	8/8
IMR90 fibroblast	SeqCap EZ	5,019,916,240	18,220	92%	—	—

Quality-filtered sequence represents the total amount of sequence data generated that passed the Illumina GA IIx quality filter (bp, base pair). The number of high-quality coding variants is the number of variants found with a sequencing depth of at least eight and a consensus quality score of at least 30. The dbSNP percentage represents the percentage of identified variants present in the Single Nucleotide Polymorphism Database. The shared coding region is the portion of the genome, in base pairs, that was sequenced at high depth and quality in both the iPS cell line and its progenitor fibroblast. The number of coding mutations lists both the number of identified coding mutations and a projection of the total number of identified mutations based on the fraction of Consensus Coding Sequence variants[16] (out of ~17,000 expected variants) successfully identified in both hiPS cells and fibroblasts.

* For these cell lines, mutation calling was performed individually using both padlock probe data and hybridization-capture data. Each method found five mutations, four of which were shared, leading to a total of six mutations. Padlock probe and hybridization capture have separate strengths (specificity versus unbiased coverage); it seems that these factors directly affect the ability to find separate mutations.

Table 3.2. Genes found to be mutated in coding regions in hiPS cells

Cell line	Mutated genes	No. of non-silent mutations	No. detectable at low frequency in fibroblasts (present/tested)
CF-RiPS1.4	*OR52E8*, *TEAD4*	1	NA
CF-RiPS1.9	*OR52E8*, *FAM171A1*, *TMED9*, *TEAD4*, *RASEF*	3	NA
CV-hiPS-B	*MMP26*, *DYNC1H1*, *VMO1*, *DSC3*, *CELSR1*, *FLT4*, *UBE2CBP*, *ARHGEF5*, *IGF2BP3*, *DLG3*	7	7/8
CV-hiPS-F	*IQGAP3*, *SPEN*, *TNR*, *PBLD*, *OR6Q1*, *INTS4*, *GSG1*, *NTRK3*, *DNAH3*, *GOLGA4*, *FAT2*, *C6orf25*, *UBR5*, *SDR16C5*	12	4/7
DF19.11	*SPATA21*, *RGS8*, *LPPR4*, *KCNJ8*, *SETBP1*, *ZNF471*, *TMEM40*	5	NA
DF6-9-9	*ZZZ3*, *AKR1C4*, *NEK5*, *DAPL1*, *ITCH*, *PPP1R2*	5	0/5
dH1CF16-iPS1	*IRGQ*, *TM9SF4*	1	NA
dH1CF16-iPS4	*PKP1*, *MYOG*, *ABCA3*, *PTPRM*, *RANBP3L*, *CALN1*	4	NA
dH1F-iPS8	*CABC1* (*ADCK3*), *C1orf100*, *OR5AN1*, *CACNG3*, *MYRIP*, *SLC1A3*, *DSP*, *KLRG2*	6	NA
dH1F-iPS9	*SEMA6C*, *MYRIP*, *SLC1A3*	3	NA
FiPS3F1	*SORCS3*, *GLRA3*, *CARM1*, *EPB41L1*	2	NA
FiPS4F7	*GDF3*, *ZER1*	2	NA
iPS11a	*GTF3C1*, *SALL1*, *SLC26A3*, *ZNF16*	3	1/1
iPS11b	*MARCKSL1*, *PRDM16*, *ATM*, *LRP4*, *TCF12*, *SH3PX3* (*SNX33*), *OSBPL3*	5	0/1
iPS17a	*HK1*, *ANKRD12*, *SCN1A*, *IFNGR1*	4	NA
iPS17b	*HK1*, *CCKBR*, *ANKRD12*, *SCN1A*, *IFT122*	5	1/1
iPS29A	*PRICKLE1*, *RFX6*	2	2/2
iPS29e	*C14orf174* (*SAMD15*), *NTRK3*, *VAC14*, *ASB3*, *STX7*, *POLR1C*, *LINGO2*	6	1/4
iPS4.7	*POLE*, *UBA2*, *L3MBTL2*, *C4orf41*	2	NA
PGP1-iPS	*C11orf67*, *OSBPL8*, *NEK11*	1	1/3
FiPS4F2	*TMEM57*, *RANBP6*, *CTSL1*, *SAV1*, *KRT25*, *BCL2L12*, *LGALS1*, *TTYH2**, *COPA**, *ARSB**, *MT1B**	7	NA
FiPS4F-shpRB4.5	*NTRK1*, *CD1B*, *LRCH3*, *SH3TC1*, *GPC2*, *CDK5RAP2*, *MYH4*, *TRMU*	5	NA

The full details of each mutation are in Supplementary Table 1.
* Mutation was observed at passage 40 but not at passage 9. FiPS4F2 was sequenced at both passage 9 and passage 40. Seven mutations were present after reprogramming (FiPS4F2P9), and four more became fixed after extended culturing (FiPS4F2P40). All seven mutations found after reprogramming were also present after extended culturing.

Supplementary Table 3.1. Digital Quantification results. 17 of the 32 tested mutations were detected at low levels in progenitor fibroblasts.

Cell Line	Mutation	Gene Name	Distinguishable from Sequencing Error?	Frequency in Fibroblasts (Lower Bound, 99% Conf.)	Frequency in Fibroblasts (Upper Bound, 99% Conf.)
CV-Fibroblast	chr3,37344209,G/A	GOLGA4	Yes (Z = 7.24)	4.4 in 10,000	5.2 in 10,000
CV-Fibroblast	chr12,13132112,C/T	GSG1	Yes (Z = 2.82)	4.3 in 10,000	5.1 in 10,000
CV-Fibroblast	chr15,86323665,A/C	NTRK3	No (Z = -1.48)	0	Less than 6 in 1,000,000
CV-Fibroblast	chr11,57555913,C/G	OR6Q1	Yes (Z = 7.17)	3.1 in 10,000	3.6 in 10,000
CV-Fibroblast	chr10,69721945,G/A	PBLD	No (Z = -0.78)	0	Less than 2 in 10,000
CV-Fibroblast	chr8,57387320,C/A	SDR16C5	Yes (Z = 3.25)	4.2 in 10,000	4.9 in 10,000
CV-Fibroblast	chr1,16128108,T/A	SPEN	No (Z = -0.43)	0	Less than 1 in 100,000
CV-Fibroblast	chr18,26842003,1,A/T	DSC3	Yes (Z = 7.89)	2.2 in 10,000	3.2 in 10,000
CV-Fibroblast	chr17,4635976,1,C/G	VMO1	Yes (Z = 11.98)	7.3 in 10,000	9.7 in 10,000
CV-Fibroblast	chr7,23348208,1,C/T	IGF2BP3	Yes (Z = 4.81)	2.0 in 10,000	3.1 in 10,000
CV-Fibroblast	chr6,83659475,1,G/T	UBE2CBP	No (Z = -2.29)	0	Less than 7 in 100,000
CV-Fibroblast	chr14,101584690,1,G/A	DYNC1H1	Yes (Z = 9.57)	7.7 in 10,000	9.6 in 10,000
CV-Fibroblast	chr7,143707948,1,T/C	ARHGEF5	Yes (Z = 6.35)	5.4 in 10,000	7.0 in 10,000
CV-Fibroblast	chr11,4967620,1,A/G	MMP26	Yes (Z = 6.78)	2.9 in 10,000	4.0 in 10,000
CV-Fibroblast	chr5,179989634,1,C/T	FLT4	Yes (Z = 6.68)	3.0 in 10,000	4.4 in 10,000
Foreskin Fibroblast	chr2,159360184,T/G	DAPL1	No (Z = -0.48)	0	Less than 2 in 100,000
Foreskin Fibroblast	chr20,32522955,A/G	ITCH	No (Z = 1.15)	0	Less than 7 in 100,000
Foreskin Fibroblast	chr13,51599552,C/T	NEK5	No (Z = -1.94)	0	Less than 4 in 100,000
Foreskin Fibroblast	chr3,196732924,G/C	PPP1R2	No (Z = 0.70)	0	Less than 7 in 100,000
Foreskin Fibroblast	chr1,77817102,T/C	ZZZ3	No (Z = -6.40)	0	Less than 5 in 10,000
PGP1 Fibroblast	chr12,75291303,1,C/G	OSBPL8	No (Z = 2.12)	0	Less than 1 in 10,000
PGP1 Fibroblast	chr11,77231220,1,A/G	C11orf67	Yes (Z = 233.96)	9.9 in 100	1.0 in 10
PGP1 Fibroblast	chr3,132364052,1,G/T	NEK11	No (Z = -3.11)	0	Less than 7 in 100,000
Hib11 Fibroblast	chr15,73729171,1,G/C	SH3PX3	No (Z = 0.83)	0	Less than 5 in 100,000
Hib11 Fibroblast	chr16,27411194,1,A/T	GTF3C1	Yes (Z = 167.93)	4.3 in 100	4.4 in 100
Hib17 Fibroblast	chr3,130708070,1,G/T	IFT122	Yes (Z = 2.73)	5.4 in 100,000	9.8 in 100,000
Hib29 Fibroblast	chr16,69288565,1,C/T	VAC14	No (Z = -10.83)	0	Less than 1 in 10,000
Hib29 Fibroblast	chr6,117355141,1,C/G	RFX6	Yes (Z = 4.72)	4.0 in 100,000	7.1 in 100,000
Hib29 Fibroblast	chr6,43596675,1,C/G	POLR1C	No (Z = -0.86)	0	Less than 3 in 100,000
Hib29 Fibroblast	chr12,41145031,1,A/G	PRICKLE1	Yes (Z = 32.60)	16 in 10,000	18 in 10,000
Hib29 Fibroblast	chr2,53831544,1,C/T	ASB3	Yes (Z = 35.37)	16 in 10,000	18 in 10,000
Hib29 Fibroblast	chr6,132823653,1,G/T	STX7	No (Z = -2.51)	0	Less than 5 in 100,000

.

Chapter 4: Functional Consequences of Coding Mutations in Induced Pluripotent Stem Cells

4.1 Abstract

Recent studies indicate that human-induced pluripotent stem cells contain genomic structural variations and point mutations in coding regions. However, these studies have focused on fibroblast-derived human induced pluripotent stem cells, and it is currently unknown whether the use of alternative somatic cell sources with varying reprogramming efficiencies would result in different levels of genetic alterations. Here we characterize the genomic integrity of eight human induced pluripotent stem cell lines derived from five different non-fibroblast somatic cell types. We show that protein-coding mutations are a general feature of the human induced pluripotent stem cell state and are independent of somatic cell source. Furthermore, we analyze a total of 17 point mutations found in human induced pluripotent stem cells and demonstrate that they do not generally facilitate the acquisition of pluripotency and thus are not likely to provide a selective advantage for reprogramming.

4.2 Introduction

The induction of pluripotency in human somatic cells by defined transcription factors represents a breakthrough in regenerative medicine[44, 55, 80-82]. The generation of patient-specific human induced pluripotent stem cells (hiPSCs) and their autologous cell derivatives would help to overcome the problems of immune rejection

and tissue availability. However, the applications of cell therapies in human patients are subject to very stringent safety requirements, and there is a general concern in the field about the safety of hiPSCs.

Successful generation of hiPSCs depends on the complete reprogramming of the somatic epigenome to a pluripotent state while the genome remains unchanged. Although initial reports demonstrated that human embryonic stem cells (hESCs) and hiPSCs were very similar, recent reports have uncovered striking genetic and epigenetic differences between these two pluripotent cell types[26, 36, 45, 83, 84]. It has been shown that hiPSCs display protein-coding mutations, large-scale genomic rearrangements, persistent epigenetic marks from the somatic cell type of origin and aberrant methylation patterns[26, 36, 85]. These findings indicated that hiPSCs contain genomic defects that could preclude their use in stem cell therapies. However, most of these studies focused on fibroblast-derived hiPSCs, and a more comprehensive analysis is essential to determine whether there are specific somatic cell types that may reprogram into hiPSCs with fewer (or perhaps none) of these aberrations. Additionally, it is unclear whether the protein-coding mutations found in hiPSCs provide any functional advantage and, thus, are selected for during the process of reprogramming.

In this work, we characterize at single-nucleotide resolution the genomic integrity of eight hiPSC lines derived from five different non-fibroblast somatic cell types with varied reprogramming efficiencies. Moreover, we functionally characterize the role of 17 point mutations found in hiPSCs for their ability to increase reprogramming efficiency. We demonstrate that the majority of these mutations do not favor the reprogramming process and suggest that most of them originated randomly or were initially present in the somatic population of origin. Our

observations of the genetic abnormalities of hiPSCs will contribute to a deeper understanding of the reprogramming process.

4.3 Methods

4.3.1 Cell culture.

The hiPSC lines ASThiPS4F4, ASThiPS4F5, HUVhiPS4F1, HUVhiPS4F3, FhiPS4F7, NSChiPS2F and FhiPS3F1 have been previously described[36, 86-88], and were obtained from existing cultures. The hiPSC lines MSChiPS4F4, MSChiPS4F8 and KhiPS4F8 met all requirements (morphology, pluripotent gene expression, normal karyotype and in vivo differentiation by teratoma formation) to define them as pluripotent. Derived hiPSCs were cultured as described[51]. 293T cells and BJ human fibroblasts (ATCC, CRL-2522) were cultured in DMEM (Invitrogen) supplemented with 10% FBS and 0.1 mM non-essential amino acids.

HUVEC cells were obtained from Lonza (C-2519A) and grown with EGM-2 media (Lonza) as recommended. MSCs were kindly provided by Ce´cile Volle (Sanofi-Aventis) and grown in a-MEM (Invitrogen) containing 10% FBS (Hyclone), penicillin/streptomycin, sodium pyruvate, non-essential amino acids, and L-glutamine (all from Invitrogen). Human keratinocytes were obtained and cultured as previously described[62].

4.3.2 hiPSC generation.

To generate hiPSCs (KhiPS4F8, MSChiPS4F4 and MSChiPS4) and to evaluate reprogramming efficiency, experiments were performed as described with minor modifications[51]. Briefly, BJ fibroblasts, keratinocytes, MSCs or HUVEC cells were infected with an equal ratio of retroviruses or retroviruses plus lentiviruses by

spinfection of the cells at 1850 r.p.m. for 1 h at room temperature in the presence of polybrene (4 mg/ml). After one (in case of the HUVEC cells), two (in case of the BJ fibroblasts or keratinocytes) or three (in case of the MSCs) viral infections, cells were trypsinized and transferred onto fresh irradiated mouse embryonic or human fibroblasts. One day after, cells were switched to hES cell medium (DMEM/F12 or KO-DMEM (Invitrogen) supplemented with 20% knockout serum replacement (Invitrogen), 1 mM L-glutamine, 0.1 mM non-essential amino acids, 55 mM b-mercaptoethanol and 10 ng/ml bFGF (Joint Protein Central)). Depending on the cell type of origin, colonies were stained for Nanog expression at day 18 (in the case of HUVEC-derived hiPS cells) or day 24 (in the case of BJ fibroblasts-derived hiPS cells), or isolated to establish cell lines.

4.3.3 Plasmid construction.

The reprogramming plasmids pMX-OCT4, pMX-SOX2, pMX-KLF4, and pMX-cMyc together with pLVTHM were obtained from Addgene (plasmids 17217, 17218, 17219, 17220 and 12247, respectively). For the construction of pMX-NTRK3, pMX-FAIM3, pMX-POLR1C, pMX-GDF3, and pMX-HK1 (fragment corresponding to the nucleotides 277–2753), specific coding region sequences were amplified by PCR from Human ORFeome library plasmids containing the corresponding cDNAs. cDNA fragments were digested with adequate restriction enzymes, purified, and subcloned into linearized pMX plasmid. For the construction of pMX-CCKBR, pMX-SAMD3, pMX-UBA2, pMX-TRAF6, pMX-MARCKSL1, pMX-CD1B, pMX-GSG1, pMX-NRP1, pMX-NEK11, pMX-CTSL1, pMX-ASB3 and pMX-ZNF16, specific pDONR223 plasmids from the Human ORFeome library containing the corresponding cDNAs were used to transfer the cDNAs to the vector pMX-GW (Addgene, 18656). The transfer was achieved by using the Gateway LR Clonase enzyme mix (Invitrogen).

The plasmids pMX-p16, pMX-CDK4, pMX-CycD1, pLVTHM-CycE and pLVTHM-p53 were generated as described[48, 51]. The plasmid pMX-RFP was kindly provided by Dr Guanghui Liu (Gene Expression Laboratory, The SALK Institute, La Jolla, CA). For the introduction of specific point mutations in the coding sequences of the above genes (see Supplementary Table 4.2 for specific mutations) the QuickChange Site-Directed Mutagenesis kit was used (Stratagene). For the generation of plasmids encoding shRNAs against the genes used in this study, specific oligonucleotides were annealed, phosphorylated with T4 kinase and ligated into MluI/ClaI-linearized pLVTHM plasmid. The design of three different pairs of shRNAs was carried out using the SFold software (http://sfold.wadsworth.org/), and knockdown efficiency was assayed in 293T cells. The most efficient pairs of shRNAs were assayed in HUVEC or BJ fibroblasts cell and used in the corresponding experiments. All constructs generated were subjected to direct sequencing to rule out the presence of mutations.

4.3.4 Retroviral and lentiviral production.

Moloney-based retroviral vectors (pMX and derived) and second-generation lentiviral vectors (pLVTHM and derived) were co-transfected with packaging plasmids to generate viral particles in 293T cells using Lipofectamine (Invitrogen) as previously described[51].

4.3.5 Immunostaining.

Immunofluorescence analysis for the detection of pluripotent markers in hiPSCs or for the detection of differentiation-associated markers in teratomas was performed as described[88]. Immunohistochemical/immunofluorescent detection of Nanog and Tra-1-60 was performed as described[51].

4.3.6 RNA isolation and real-time PCR analysis.

Total RNA was isolated using Trizol Reagent (Invitrogen) according to the manufacturer's recommendations. cDNA was synthesized using the SuperScript II Reverse Transcriptase kit for RT–PCR (Invitrogen) or the RT Supermix M-MuLV kit (BioPioneer). Real-time PCR was performed using the SYBR-Green PCR Master mix (Applied Biosystems) in the ViiA 7 Real Time PCR System (Applied Biosystems). Glyceraldehyde 3-phosphate dehydrogenase expression was used to normalize values of gene expression and data is shown as fold change relative to the value of the sample control. All the samples were done in triplicate.

4.3.7 Whole-genome library construction.

Library construction was performed as previously described[36, 89]. Briefly, for each sample, roughly 1.5–3 mg of genomic DNA (in 100 ml volumes) was sheared with a Covaris AFA. The fragmented genomic DNA was end repaired, A-tailed, and ligated to sequencing adaptors, with a purification step between each process. The purified ligated products were then amplified by PCR to generate whole-genome libraries.

4.3.8 In-solution hybridization capture with DNA baits.

Liquid exome capture was performed as previously described[36].

4.3.9 Consensus sequence generation and variant calling.

Variant calling was performed as previously described[36]. Briefly, reads obtained from the Illumina Genome Analyzer were post-processed and quality filtered using GERALD, mapped to the genome using BWA, downsampled using Picard and used to generate a consensus sequence for each sample using GATK. The consensus sequences were then compared with find candidate novel mutations in

hiPSCs[36]. Sites where each hiPSC line showed heterozygous SNPs not observed in the progenitor line were considered as candidate mutations if no allelic content was present in the somatic progenitor and if the candidate mutation had not previously been observed in other samples or the dbSNP database.

4.3.10 Sanger validation of candidate mutations.

Genomic DNA of both the hiPSC line and its somatic progenitor (6 ng each) was amplified in separate 50 ml PCR reactions with 100 nM of specifically designed forward and reverse primers around the mutation site (primers available under request) and 25 ml of Taq 2x master mix (NEB) at 94 ºC for 2min, followed by 35 cycles of 94 ºC for 30s, 57 ºC for 30s and 72 ºC for 30 s, and final extension at 72 ºC for 3 min. The PCR products were then purified with Qiagen Qiaquick columns, and 10 ng of purified DNA was pre-mixed with 25 pmol of the forward primer for Sanger sequencing at Genewiz Inc.

4.3.11 Statistical analysis/TiGER database.

To check for enrichment of reprogramming-associated mutations in genes that are expressed in a tissue-specific manner, the fraction of UniGene IDs corresponding to mutated genes called as ‘tissue-specific’ in the TiGER90 database was identified as 49/132 (37%). As 6,699/19,526 (34%) of the genes annotated in the TiGER database are considered to be tissue specific, a $\chi 2$-test with one degree of freedom can be used to test for equivalency of distribution. The obtained $\chi 2$ value is 0.460, indicating that the fraction of mutated hiPSC genes that are tissue specific is not significantly different than that found in a random sample of genes ($P = 0.4975$). Reprogramming-associated mutations therefore do not appear to be enriched in tissue-specific genes.

4.3.12 Statistical analysis/active and inactive chromatin states.

To check for enrichment of reprogramming-associated mutations in active or inactive chromatin, we utilized a χ^2-test with three degrees of freedom to test for equivalency of distribution. We identified the chromatin state of each mutated gene using previously published data[91]. These data divided each gene into one of four categories: no trimethylation, H3K4 trimethylation, H3K27 trimethylation, or both. We compared the distribution of mutated genes across each of these four categories with the expected distribution for all genes in three cell types: fibroblasts, ESCs and iPSCs[91]. The obtained χ2 values were 1.03 (P = 0.79), 3.78 (P = 0.29) and 6.97 (P = 0.07), respectively, indicating that the distribution of mutated hiPSC genes in each chromatin region is not significantly different than expected by random chance (a = 0.01). Reprogramming-associated mutations therefore do not appear to be enriched in active or inactive chromatin states.

4.3.13 Non-coding versus coding mutations.

To compare the mutation rates per base pair in coding and non-coding regions of the genome, variant calling was performed as above on non-coding regions of the genome surviving library enrichment in eight hiPSC lines and their progenitor lines. The mutation rate per base pair was then estimated by dividing the number of candidate coding and non-coding mutations by the number of exomic and non-coding base pairs covered. The average coding and non-coding mutation rates were compared.

4.4 Results

4.4.1 hiPSC lines from varied cell types contain protein-coding mutations.

We previously sequenced the protein-coding regions of 22 fibroblast-derived hiPSC lines and discovered that the hiPSCs analyzed carried between 2 and 14 point mutations in protein-coding regions[36]. In this study, we sought to determine whether low reprogramming efficiency (and therefore a potentially higher level of selection pressure that could allow the fixation of advantageous mutations) or cell type of origin (as fibroblasts could possess a higher somatic mutation rate than other cell types) could contribute to the overall reprogramming-associated mutational load. To this end, we performed targeted exome sequencing on eight non-fibroblast-derived hiPSC lines and their five somatic cell types of origin using an in-solution hybridization capture method (Supplementary Table 4.1). Somatic mutations in each hiPSC line were identified via pairwise comparison with the matched somatic cell of origin and independently confirmed with capillary Sanger sequencing. We identified a total of 40 point mutations throughout all the hiPSC lines analyzed, leading to an average of five coding mutations per line (Table 4.1). As we identified ~89% of expected total single-nucleotide polymorphisms at high sequencing depth in protein-coding regions, this led to a projection of 45 total mutations in protein-coding regions, or ~6 coding mutations per cell line. The levels of mutational load from each individual somatic cell type were statistically indistinguishable, and within the range previously observed for fibroblast-derived hiPSC lines[36] (Table 4.1). These results indicate that hiPSC-associated mutations cannot be avoided by using younger or potentially more genetically protected somatic cell sources as progenitor cells. Moreover, we determined that reprogramming efficiency, which varies between 0.001 to 3% for

these cell types, did not seem to have a measurable effect on the hiPSC mutational load. Thus, reprogramming-associated point mutations appear to be a general feature of hiPSCs.

We next investigated whether mutations in hiPSCs were either enriched or depleted in protein-coding regions. To this end, we examined additional non-coding regions captured in our sequencing analysis, and found a similar mutation rate per base pair analyzed for both coding and non-coding regions (Table 4.2). We also investigated whether point mutations in hiPSCs tended to occur in active/ubiquitous or silent/tissue-specific genes. Among a total of 132 mutated genes (from this and previous studies[36]) annotated in the TiGER Database[90], 37% of these genes showed tissue-specific expression, which is very similar to the overall level of tissue specificity observed in the genes annotated in the database (34%; P = 0.4975), indicating that mutations are not preferentially occurring in silent genes. We additionally checked for any potential enrichment of mutations in active or inactive transcriptional regions of the genome[91]. We found that mutations were not significantly enriched in the active or inactive chromatin regions of fibroblasts (P = 0.79), hESCs (P = 0.29) or hiPSCs (P = 0.07). Furthermore, only one gene (NTRK3) was found mutated in more than one independent hiPSC line, and mutated genes did not cluster in a specific functional pathway. These combined findings suggest that mutations in hiPSCs are spread throughout both transcriptionally active and silent regions of the genome.

4.4.2 hiPSC-point mutations do not favor the process of reprogramming.

We previously showed that at least half of reprogramming-associated point mutations pre-exist in starting somatic cell populations at low frequency[36]. This leads to a hypothesis that a sub-population of somatic cells carrying certain mutations could be primed for reprogramming, which would be consistent with the elite model for

reprogramming[92]. To investigate the functional potential of these mutations during reprogramming, we first assessed whether mutated alleles were expressed in the hiPSC lines. We isolated RNA from three hiPSC lines, reverse-transcribed it into cDNA and sequenced a total of six transcripts of randomly selected genes found mutated in these hiPSC lines. We detected heterozygous expression of both mutant (mut) and wild-type(wt) alleles in all cases (Fig. 4.1), indicating that mutated transcripts are expressed in hiPSCs.

We next sought to determine whether reprogramming-associated mutations could contribute functionally in facilitating the acquisition of pluripotency during reprogramming. From a total of 164 different genes found mutated in hiPSC lines[36], we assayed the function of 17 candidate genes and their mutated forms during reprogramming (Supplementary Table 4.2). These candidate genes were selected based on the likelihood of the mutation to change protein function, the mutation type (only non-synonymous mutations were analyzed) and whether the gene was known to be related to the maintenance and/or acquisition of pluripotency[36] (Table 4.1; Supplementary Fig. 4.1; Supplementary Table 4.2). We also analyzed the expression of these 17 genes in BJ fibroblasts, human umbilical vein endothelial cells (HUVEC), hESC and hiPSC lines to ensure gene expression in at least one of the somatic cell types used in this work. Owing to the difficulty in predicting the functional consequences of each specific mutation, we first performed “loss-of-function” reprogramming experiments to mimic a possible diminished activity or protein instability of the mutated form. To this end, we designed a panel of lentiviruses encoding short hairpin RNAs (shRNAs) against the selected genes, and coinfected each separately with retroviruses expressing OCT4, SOX2, KLF4 and cMyc (OSKC) in BJ fibroblasts (Fig. 4.2a). Moreover, to determine whether these effects were cell-

type specific, we performed similar reprogramming experiments in HUVEC (Supplementary Fig. 4.2a). If a genetic mutation was selected for its ability to facilitate reprogramming due to a loss of protein function, it would be expected that downregulation of the mutated gene would increase reprogramming efficiency. A decrease in reprogramming efficiency was detected after downregulation of FAIM3, SAMD3, ZNF16, MARCKSL1, NRP1, TRAF6, GSG1 and HK1, whereas no significant changes were detected after the downregulation of all but one of the assayed genes, POLR1C (Fig. 4.2a, Supplementary Fig. 4.2a, Supplementary Fig. 4.2b). Interestingly, we observed that downregulation of POLR1C in BJ fibroblasts, but not in HUVEC, resulted in an increased reprogramming efficiency. However, it is unclear whether the specific reprogramming-associated mutation in POLR1C would result in the same phenotype. Overall, our data suggest that protein-coding point mutations generally do not prime rare cells for reprogramming through the loss-of-function mechanism.

Next, we performed ‘gain-of-function’ reprogramming experiments to determine whether expression of the mutated form facilitated cell reprogramming. To this end, we designed a panel of retroviruses encoding both the wt form and the corresponding mutated form found in hiPSCs of each specific gene (see specific mutations in Supplementary Table 4.2), and coexpressed them with OSKC in BJ fibroblasts and HUVECs (Fig. 4.2b, Supplementary Fig. 4.2c). If a mutation were selected during reprogramming due to a gain-of-function, it would be expected that expression of the mutated form would increase the reprogramming efficiency. We observed that only the expression of HK1 slightly increased reprogramming efficiency (Fig. 4.2b and Supplementary Fig. 4.2c). Importantly, we did not observe significant differences in reprogramming efficiency between cells overexpressing the mutated

forms and cells overexpressing their respective wt forms (Fig. 4.2b), indicating that the presence of the mutated protein does not increase reprogramming efficiency.

We have previously shown that both the mut allele and the wt allele are expressed in hiPSCs (Fig. 4.1). However, it is possible that a similar level of expression of the wt and mut protein forms is necessary in order for the mutation to influence reprogramming efficiency in a gain-of-function manner. To clarify this, we performed a reprogramming experiment where OSKC were coexpressed together with a similar total amount of retrovirus encoding either only the wild type form or both the wt and mut forms of a mutated gene in an equal ratio (1:1). Using this strategy, we were able to compare the reprogramming efficiency of cells overexpressing wt and mutated protein (wt/mut) in equal amounts with that of cells overexpressing wt protein alone (wt/wt). Interestingly, we did not observe any difference in reprogramming efficiency between cells overexpressing the wt/wt and wt/mut proteins (Fig. 4.3a). Finally, we investigated whether silencing of retroviral transgenes during reprogramming could mask a gain-of-function effect of the mutated genes at a later stage of reprogramming. We thus analyzed the reprogramming efficiency of cells infected with retroviruses expressing OSKC, the wt or mutated forms of the genes evaluated in this study, and a red fluorescent protein (RFP) reporter gene to monitor transgene silencing. Reprogramming efficiency was evaluated based on the number of Tra-1-60þ/RFPþ colonies present at day 14. These colonies represent putative bona-fide hiPSC colonies, as they express the stem cell marker Tra-1-60 but lack silencing of the exogenous transgenes. Thus, we only considered reprogramming events where transgene expression was still active. Importantly, we did not observe differences in reprogramming efficiency between cells overexpressing the mutated forms and cells overexpressing their respective wt forms (Fig. 4.3b). Furthermore, we

also evaluated reprogramming efficiency in the same experiment at day 14 by analyzing the number of Tra-1-60 þ / RFP colonies (evaluating putative bona-fide hiPSC colonies where transgene silencing occurred), and obtained a similar result (data not shown). Overall, these data suggest that most of these mutated genes do not facilitate reprogramming through a gain-of-function or loss-of-function mechanism.

4.5 Conclusions

Our work demonstrates that hiPSCs contain protein-coding mutations independent of the cell type of origin (as we analyzed hiPSC lines derived from five tissue types). Moreover, we determined that reprogramming efficiency, and therefore the level of selection pressure which could allow the fixation of advantageous mutations, did not to have a measurable effect on the hiPSC mutational load. Although the functional consequences of individual protein-coding mutations detected in hiPSCs remain to be characterized, these alterations could potentially contribute to the functional differences observed between hiPSC lines[93-95].

Two independent groups have recently reported the whole-genome sequencing of human and murine iPSC lines and their corresponding somatic cell lines[96, 97]. They identified hundreds of single-nucleotide variants (SNVs) in non-coding regions and an average of 6–12 SNVs in coding regions[96, 97], which is consistent with our results[36]. Importantly, their data suggest that much of the genetic variation in iPSC clones pre-exists in the somatic population of origin and is fixed as a consequence of cloning individual cells during iPSC generation[96, 97]. Although these reports supported previous observations[36], they did not investigate whether identified

mutations contributed functionally to facilitate the acquisition of pluripotency during reprogramming.

In this work, we show evidence suggesting that most reprogramming-associated point mutations do not provide a detectible selective advantage towards a reprogrammed state. As inhibiting wt POLR1C expression had a positive impact on reprogramming efficiency, we cannot rule out a potential role of the mutation found in POLR1C in facilitating reprogramming. If this is the case, the fact that downregulation of POLR1C increases reprogramming efficiency in fibroblasts, but not in HUVECs, could indicate the existence of tissue-specific mutations affecting reprogramming efficiency, as POLR1C^{P278R} was found in one hiPSC line derived from human fibroblasts. Although it remains possible that untested mutated genes or a combination of mutations in a certain cellular context could have a role, the finding that only one gene (NTRK3) was found mutated in 2 out of 30 independent hiPSC lines, that mutated genes do not cluster in a specific functional pathway that could explain their selection during the reprogramming process, and that non-coding regions showed a similar mutational load, indicate that reprogramming-associated mutations seem to occur through a random process without selection and/or are initially present in the somatic population of origin[96, 97]. It has been suggested that genomic alterations (that is, duplications, deletions and mutations) are selected for during reprogramming, yet this has not been demonstrated[26, 36, 45, 83-85]. In contrast to well-established recurrent genomic aberrations (for example, chromosome 12 duplications) present in hESC or hiPSC lines that are functionally selected upon prolonged culture[45], our results suggest that reprogramming-associated point mutations generally do not generally affect reprogramming efficiency (although there could be exceptions). To our knowledge, the data provided herein provides for the

first time a functional analysis of the role of specific genomic alterations (that is, point mutations in coding regions) on the reprogramming process and have potential implications for the future of the hiPSC field in regenerative medicine.

4.6 Acknowledgements

We express our gratitude to Travis Berggren, Margaret Lutz and Veronica Modesto for their support at the Salk Institute-Stem Cell Core, to Joaquin Sebastian for critically reading the manuscript, to Guanghui Liu for sharing reagents and to the rest of the Belmonte lab. A.G. was supported by the Focht-Powell Fellowship and a CIRM predoctoral fellowship. Work in this manuscript was supported by grants from Fundacion Cellex, TERCEL-ISCIII-MINECO, Sanofi, National Institutes of Health and the G. Harold and Leila Y. Mathers Charitable Foundation.

Chapter 4, in part, is a reprint of the material as it appears in: Sergio Ruiz*, Athurva Gore*, Zhe Li, Athanasia D. Panopoulos, Nuria Montserrat, Ho-Lim Fung, Alessandra Giorgetti, Josipa Bilic, Erika M. Batchelder, Holm Zaehres, Hans R. Scholer, Kun Zhang, and Juan Carlos Izpisua Belmonte. “Analysis of protein-coding mutations in hiPSCs and their possible role during somatic cell reprogramming.” *Nature Communications*. 2013 January 22; 4: 1382. doi:10.1038/ncomms2381. Used with permission. The dissertation author was one of the primary investigators and authors of this paper.

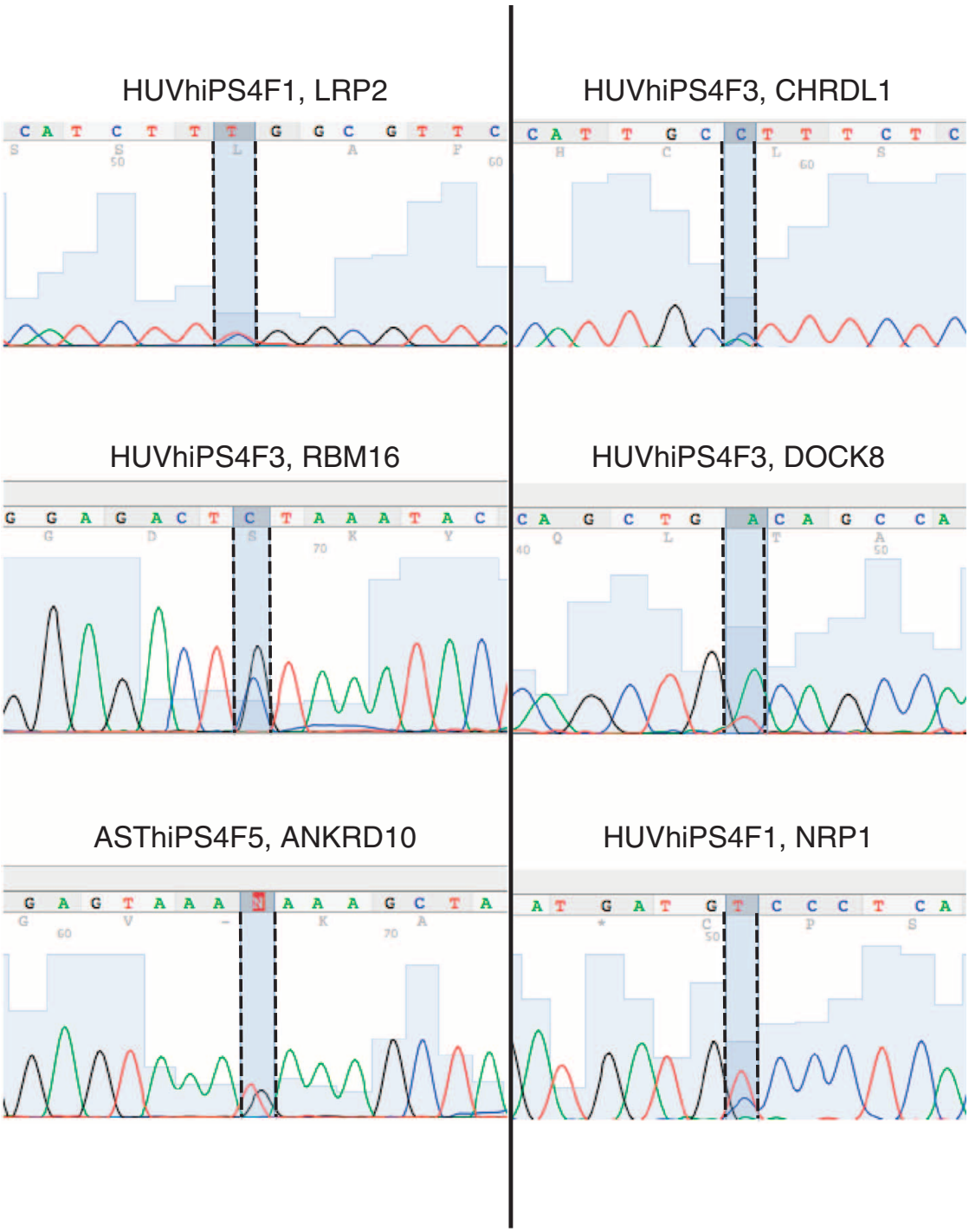


Figure 4.1. Mutated alleles are expressed in hiPSC lines. Sanger chromatograms showing the results of RNA Sequencing analysis performed on the indicated genes found mutated in the indicated hiPSC lines. Dashed lines highlight the point-mutated nucleotide. Note the expression of both reference and mutated alleles in all cases analyzed.

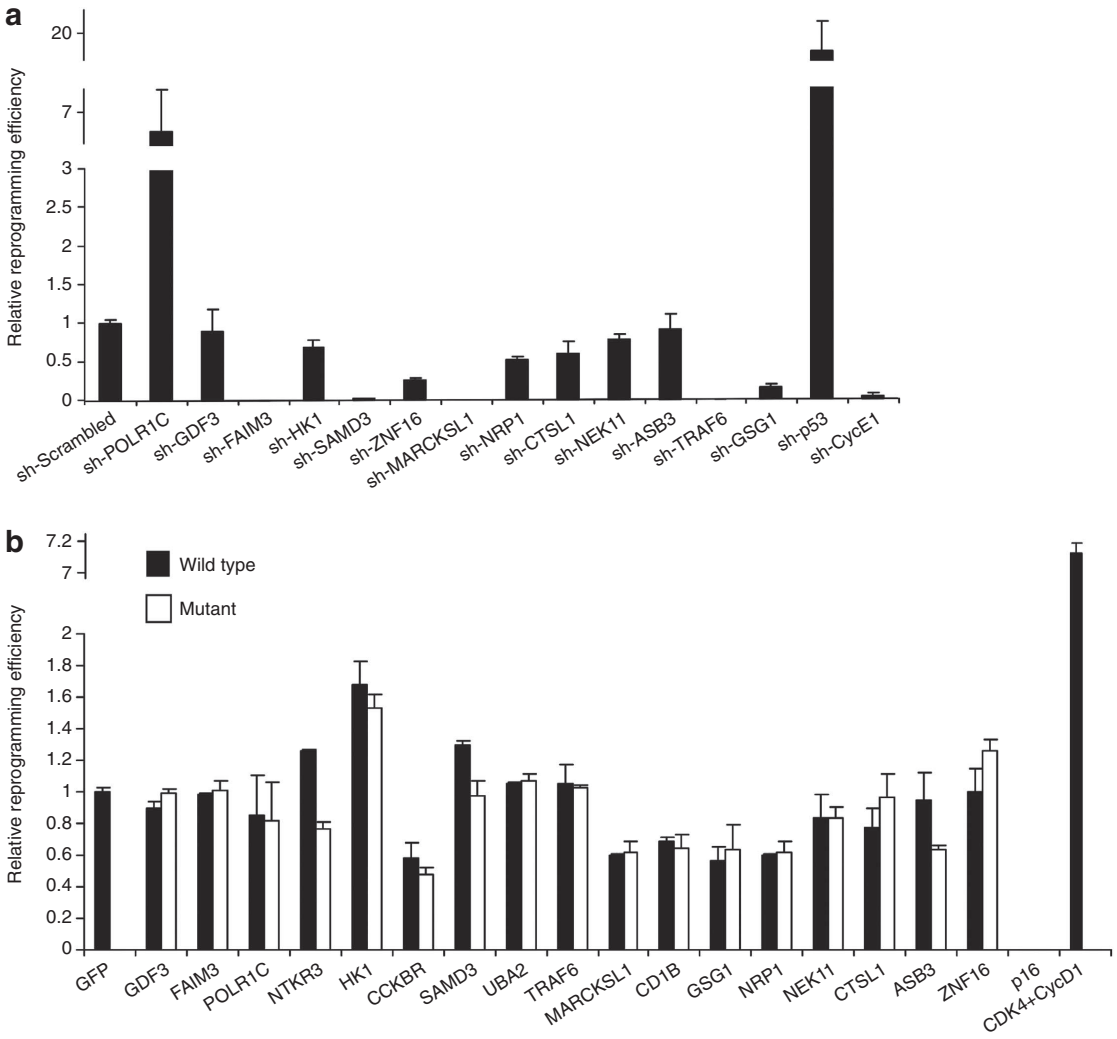


Figure 4.2. Evaluation of the functional effect of hiPSC mutations on reprogramming efficiency. (a,b) Human BJ fibroblasts were infected with retroviruses encoding OSKC, and either lentiviruses encoding shRNAs against the indicated proteins (a) or retroviruses encoding the wild type or mutated proteins (b). Relative reprogramming efficiencies (evaluated as percentage of Nanog þ colonies) are shown as fold change normalized to the averaged efficiency observed in (a) pLVTHM or (b) pMX–GFP-infected fibroblasts. In a, lentiviruses encoding shRNAs against CycE1 or p53 were used as controls of reduced or increased reprogramming efficiency, respectively. In (b), retroviruses encoding p16 or the pair CDK4/CycD1 were used as controls of reduced or increased reprogramming efficiency, respectively. For (a), 20,000 infected cells were plated when shRNAs against POLR1C and p53 were used, and 70,000 infected cells were plated under all other conditions. For (b), a total of 25,000 infected cells were plated under all conditions. Two independent experiments with two biological replicates were carried out. All error bars depict the s.d.

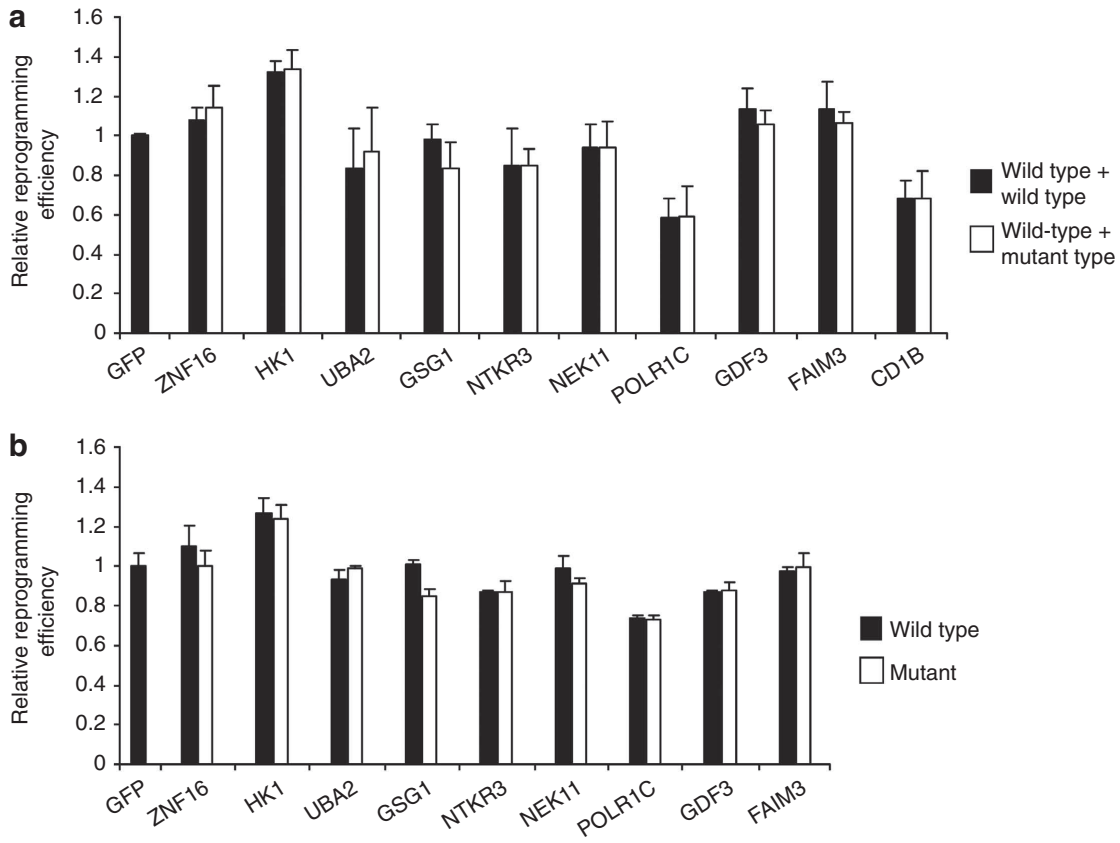
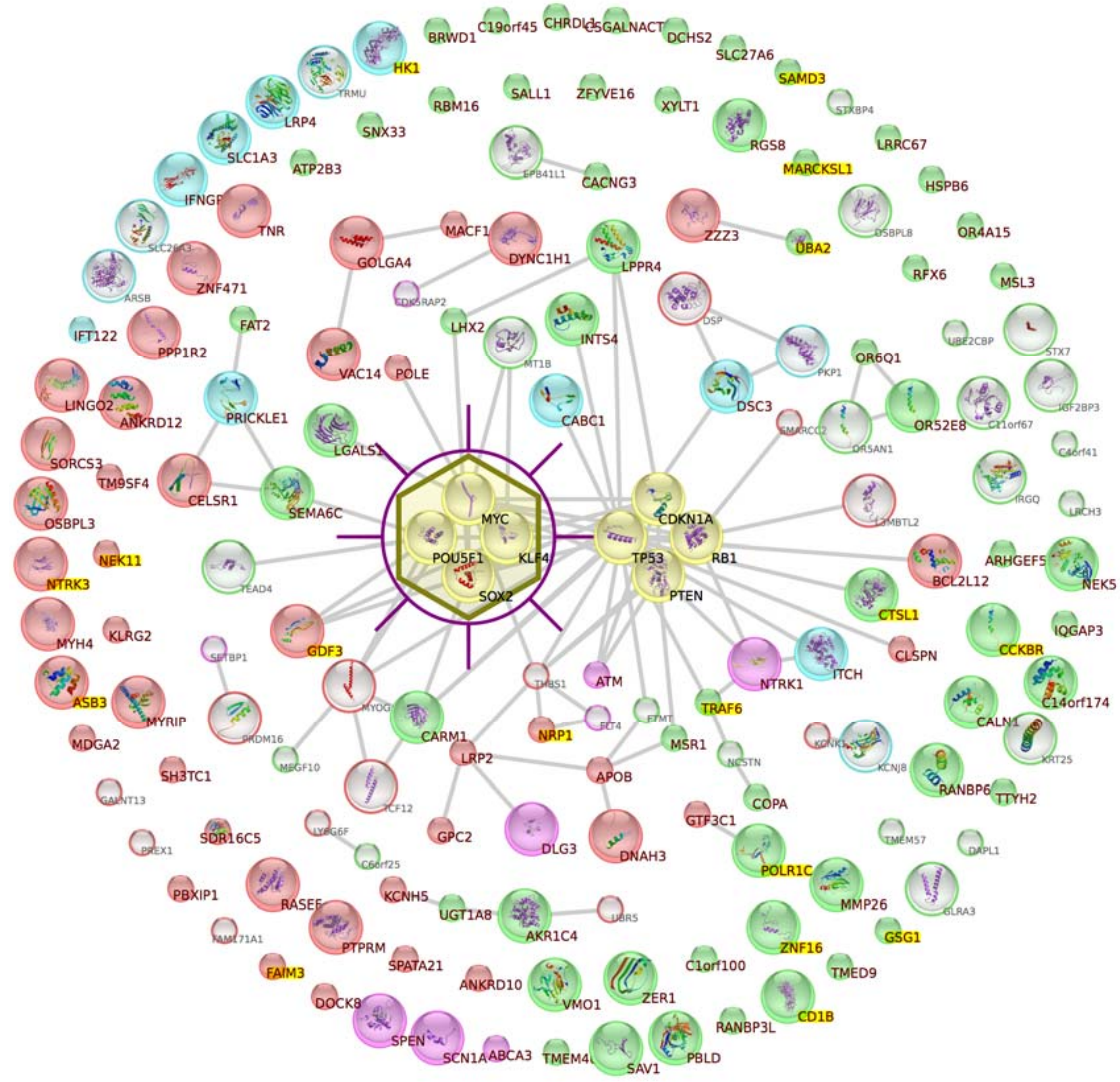
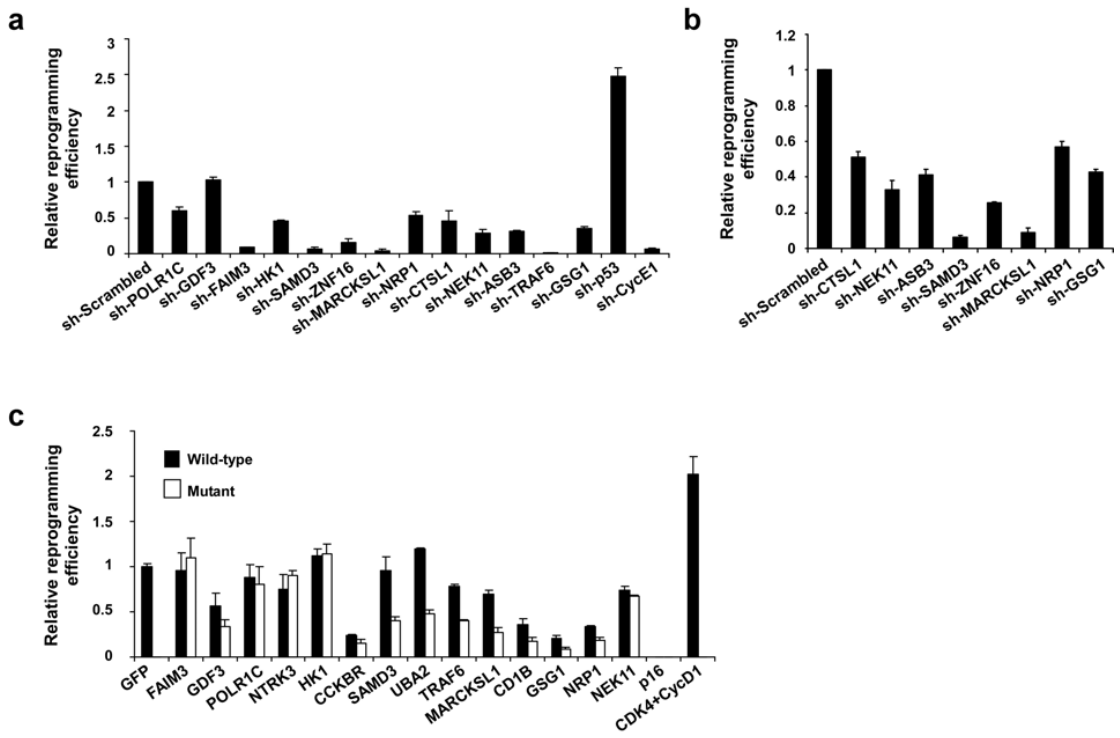


Figure 4.3. Retroviral silencing and wild-type/mutant gene ratios do not alter reprogramming efficiency. (a) HUVEC cells were infected with retroviruses encoding OSKC, and a similar total amount of retroviruses encoding only the wild-type form or both, the wild-type and mutant forms of the protein in an equal proportion. (b) HUVEC cells were infected with retroviruses encoding OSKC, RFP and the wild-type or mutated forms of the genes indicated. Relative reprogramming efficiencies (evaluated as percentage of Tra-1-60 þ colonies) are shown as fold change normalized to the averaged efficiency observed in green fluorescent protein-infected HUVECs. Ten thousand infected cells were plated under all the conditions. Two independent experiments with two biological replicates were carried out. All error bars depict the s.d.



Supplementary Figure 4.1. Some functionally tested genes are related to reprogramming factors or common cancer genes. Mutated genes found in iPSC lines from this study and Gore et al, 2011 are depicted. Each mutated gene is represented by a colored sphere (with red signifying a gene known to be mutated in cancer by COSMIC v49 and blue signifying a gene related to Mendelian disorders), and each protein-protein interaction is represented by a connecting line. Interactions between each mutated gene, the OSKC reprogramming factors, and four common cancer genes (PTEN, P53, P21, and RB1) are shown. The highlighted genes were functionally tested. Several genes with known interactions with other mutated genes, reprogramming factors, or cancer genes were chosen for functional testing. The figure was made using the Cytoscape network visualization program and the Inkscape drawing program. Protein-protein interactions were taken from the String database.



Supplementary Figure 4.2. Evaluation of the functional effect of mutations found in hiPSC lines on reprogramming efficiency. (a-c) HUVEC cells were infected with retroviruses encoding OSKC, and either lentiviruses encoding shRNAs against the indicated proteins (a, b) or retroviruses encoding the wild type or mutated proteins (c). Relative reprogramming efficiencies evaluated as percentage of Nanog+ (a, c) or Tra-1-60+ colonies (b) are shown as fold change normalized to the averaged efficiency observed in pLVTHM (a, b) or pMX-GFP (c) infected HUVECs correspondingly. In (a) lentiviruses encoding shRNAs against CycE1 or p53 were used as controls of reduced or increased reprogramming efficiency. In (b) retroviruses encoding p16 or the pair CDK4/CycD1 were used as controls of reduced or increased reprogramming efficiency, respectively. For (a, b), 10,000 infected cells were plated under all the conditions. For (c), a total of 12,000 infected cells were plated under all conditions. Two independent experiments with two biological replicates were carried out. All error bars depict the standard deviation.

Table 4.1. List of protein-coding mutations in hiPSC lines.

Sample	Chromosome	Position	Gene	Alleles	Protein change	Mutation type	SIFT functional prediction	Mutated in cancer?
ASThiPS4F4	6	31783527	*LY6G6F*	GAC-GAt	D122D	Synonymous	NA	Yes
ASThiPS4F4	8	68087821	*LRRC67*	CTT-aTT	L121I	Non-synonymous	TOLERATED	No
ASThiPS4F5	11	54891946	*OR4A15*	CTG-CcG	L4P	Non-synonymous	DAMAGING	No
ASThiPS4F5	13	110343392	*ANKRD10*	AAG-AAt	K225N	Non-synonymous	TOLERATED	Yes
KhiPS4F8	1	205153901	*FAIM3*	TTC-aTC	F67I	Non-synonymous	DAMAGING	Yes
KhiPS4F8	5	121215932	*FTMT*	CAC-CAt	H125H	Synonymous	NA	Yes
KhiPS4F8	14	62486817	*KCNH5*	GAC-aAC	D386N	Non-synonymous	TOLERATED	No
NSChiPS2F	5	79774746	*ZFYVE16*	TCT-TaT	S823Y	Non-synonymous	DAMAGING	No
NSChiPS2F	12	54853783	*SMARCC2*	CCA-CCg	P538P	Synonymous	NA	Yes
HUVhiPS4F1	2	169809670	*LRP2*	TCG-TtG	S1070L	Non-synonymous	TOLERATED	Yes
HUVhiPS4F1	10	33542444	*NRP1*	GGC-GaC	G497D	Non-synonymous	DAMAGING	No
HUVhiPS4F1	16	17139792	*XYLT1*	AAG-AgG	K562R	Non-synonymous	TOLERATED	No
HUVhiPS4F1	4	155376303	*DCHS2*	GGA-GtA	G2529V	Non-synonymous	DAMAGING	No
HUVhiPS4F3	6	155183150	*RBM16*	GTA-cTA	V595L	Non-synonymous	TOLERATED	No
HUVhiPS4F3	9	394921	*DOCK8*	TCA-aCA	S1012T	Non-synonymous	TOLERATED	Yes
HUVhiPS4F3	X	109889590	*CHRDL1*	CTT-aTT	L86I	Non-synonymous	TOLERATED	No
HUVhiPS4F3	19	7475243	*C19orf45*	TCA-TaA	S229*	Nonsense	NA	Yes
MSChiPS4F4	1	35998698	*CLSPN*	GTG-tTG	V471L	Non-synonymous	TOLERATED	Yes
MSChiPS4F4	1	153185686	*PBXIP1*	GAC-GgC	D363G	Non-synonymous	TOLERATED	Yes
MSChiPS4F4	2	154960801	*GALNT13*	GAA-GAg	E403E	Synonymous	NA	Yes
MSChiPS4F4	5	126704124	*MEGF10*	GTC-GTg	V74V	Synonymous	NA	Yes
MSChiPS4F4	6	130572400	*SAMD3*	ATG-tTG	M106L	Non-synonymous	DAMAGING	Yes
MSChiPS4F4	10	42974297	*CSGALNACT2*	ATG-gTG	M264V	Non-synonymous	TOLERATED	Yes
MSChiPS4F4	11	36473107	*TRAF6*	GAA-aAA	E225K	Non-synonymous	DAMAGING	Yes
MSChiPS4F4	17	50475673	*STXBP4*	GTA-GTg	V236V	Synonymous	NA	Yes
MSChiPS4F4	19	40938581	*HSPB6*	TCGCCG-TCatCG	S84S P85S	Synonymous Non-synonymous	NA DAMAGING	No
MSChiPS4F4	20	46706999	*PREX1*	GCC-GCt	A703A	Synonymous	NA	Yes
MSChiPS4F4	21	39493296	*BRWD1*	AAA-AtA	K1639I	Non-synonymous	DAMAGING	Yes
MSChiPS4F4	X	11688927	*MSL3*	TCT-TtT	S111F	Non-synonymous	DAMAGING	Yes
MSChiPS4F8	1	39703363	*MACF1*	GGC-tGC	G5698C	Non-synonymous	DAMAGING	Yes
MSChiPS4F8	1	158594563	*NCSTN*	TTG-cTG	L670L	Synonymous	NA	No
MSChiPS4F8	1	231873806	*KCNK1*	GAC-GAt	D224D	Synonymous	NA	Yes
MSChiPS4F8	2	21087987	*APOB*	CAC-CgC	H1753R	Non-synonymous	DAMAGING	Yes
MSChiPS4F8	2	234287122	*UGT1A8*	GTC-GaC	V249D	Non-synonymous	NA	Yes
MSChiPS4F8	5	128390915	*SLC27A6*	GAC-GAa	D482E	Non-synonymous	DAMAGING	No
MSChiPS4F8	8	16079769	*MSR1*	CCG-tCG	P34S	Non-synonymous	TOLERATED	Yes
MSChiPS4F8	9	125834763	*LHX2*	GAG-tAG	E393*	Non-synonymous	NA	No
MSChiPS4F8	14	46496570	*MDGA2*	TTG-aTG	L318M	Non-synonymous	TOLERATED	Yes
MSChiPS4F8	15	37669438	*THBS1*	TGC-TGt	C689C	Synonymous	NA	Yes
MSChiPS4F8	X	152498688	*ATP2B3*	TCC-TaC	S1134Y	Non-synonymous	DAMAGING	Yes

NA, not applicable; SIFT, sorting intolerant from tolerant.
*Stop codon.

Table 4.2. List of candidate non-coding mutations in hiPSC lines.

Cell line	Non-coding mutations			Exon mutation rate (per bp)	Non-exon mutation rate (per bp)
	Chromosome	Position	Mutation		
	9	111225067	C – > T		
ASThiPS4F4	11	64089233	G – > T	8.0E – 08	6.2E – 08
	13	38444609	C – > T		
	2	114429763	A – > T		
ASThiPS4F5	12	55133583	G – > T	8.0E – 08	1.0E – 07
	16	2290223	G – > T		
	17	40078501	C – > T		
	5	149190453	C – > A		
	9	5175241	C – > T		
FiPS3F1	10	45274877	G – > T	1.6E – 07	1.6E – 07
	11	85134161	T – > C		
	19	48465587	C – > A		
	1	171784008	C – > A		
FiPS4F7	2	116251932	C – > A	1.2E – 07	1.3E – 07
	2	189575154	C – > A		
	9	98839743	G – > A		
HUVhiPS4F1	11	17069813	G – > A	1.6E – 07	7.2E – 08
	19	21056778	G – > A		
	2	102515666	G – > T		
HUVhiPS4F3	11	12908191	G – > T	1.6E – 07	1.1E – 07
	15	25902050	G – > A		
	22	18130926	C – > T		
	5	96143123	C – > T		
KhiPS4F8	9	122778753	A – > T	1.2E – 07	1.8E – 07
	10	85962168	C – > G		
	17	71457091	T – > C		
	5	110120401	G – > T		
	9	115077723	C – > T		
NSChiPS2F	9	127398270	C – > T	8.0E – 08	1.2E – 07
	11	9457743	T – > A		
	19	46045755	T – > G		
	1	6092988	C – > G		
	2	88885962	C – > T		
	2	230820856	T – > C		
	3	51088015	G – > T		
	4	67994660	T – > A		
	4	156930655	C – > A		
	5	156683693	G – > C		
MSChiPS4F4	6	73887041	C – > T	4.8E – 07	4.2E – 07
	6	129865735	A – > G		
	8	24379410	T – > A		
	10	94807938	T – > C		
	10	100179495	G – > A		
	14	104534424	G – > A		
	15	61826051	G – > A		
	18	9551942	C – > T		
	X	70543403	T – > C		
	X	152791107	A – > T		
	1	39703363	G – > T		
	1	46422702	G – > C		
	1	74608556	C – > A		
	1	85589993	A – > G		
	2	88885962	C – > T		
	2	128062632	A – > T		
	3	12183849	G – > A		
	4	3200574	C – > T		
	4	95415046	A – > G		
	4	144675426	G – > T		
	6	37430075	G – > T		
MSChiPS4F8	6	90653308	G – > T	4.4E – 07	6.5E – 07
	6	134570140	T – > C		
	7	37123316	T – > C		
	7	117661660	A – > C		
	9	15258029	C – > G		
	9	132317600	C – > G		
	14	29171707	C – > A		
	14	73445929	C – > G		
	16	20968574	C – > T		
	16	34177914	C – > T		
	17	7778980	G – > A		
	17	62783815	G – > A		
	18	6877390	C – > T		
	19	43784039	C – > A		
	X	138470822	G – > A		
Average				1.9E – 07	2.0E – 07

The mutation rate per base pair was similar for exonic and non-exonic regions.

Supplementary Table 4.1. Summary table for the cell lines used in this study and sequencing statistics.

Sample	Reprog. Factors	Passage number	Reprog. Efficiency	Raw Base Pairs Sequenced	Number of Variants
				Exome	
K.MMTA (human foreskin keratinocytes)		5		3,543,738,560	11,606
KhiPS4F8	Oct4, Sox2, Klf4 and cMyc	28	~ 0.1-1%	3,280,899,760	14,529
Astrocytes (human astrocytes)		12		3,890,562,960	16,495
ASThiPS4F4	Oct4, Sox2, Klf4 and cMyc	10	~ 0.1-1%	4,352,168,080	16,158
ASThiPS4F5	Oct4, Sox2, Klf4 and cMyc	10	~ 0.1-1%	4,954,615,760	16,120
HUVEC (human umbilical vein endothelial cells)		2		3,680,376,560	20,160
HUVhiPS4F1	Oct4, Sox2, Klf4 and cMyc	10	~ 2.5-3%	2,592,254,960	15,912
HUVhiPS4F3	Oct4, Sox2, Klf4 and cMyc	15	~ 2.5-3%	5,021,900,880	15,833
hNSC (human neural stem cells)				4,249,629,120	16,080
NSChiPS2F	Oct4 and Klf4	23	~ 0.006%	4,496,763,040	15,936
MSC (mesenchymal stem cells)		7		5,602,732,160	15,757
MSCiPS4	Oct4, Sox2, Klf4 and cMyc	14	N.D.	5,926,412,320	16,096
MSCiPS8	Oct4, Sox2, Klf4 and cMyc	14	N.D.	6,298,703,680	16,233

N.D.= Not Determined

Supplementary Table 4.1. Summary table for the cell lines used in this study and sequencing statistics.

Gene	Functional description and/or relation with pluripotency	Protein Change	GeneSIFT Functional Prediction	Mutated in cancer?
GDF3 (Growth Differentiation Factor-3)	Gene expressed in undifferentiated hES cells which participates in the maintenance of pluripotency	I337L	DAMAGING	YES
HK1 (Hexokinase-1)	Kinase that phosphorylates glucose to produce glucose-6-phosphate and promotes the glycolytic pathway which has been shown to facilitate cell reprogramming	I558T	TOLERATED	NO
FAIM3 (Fas apoptotic inhibitory molecule-3)	Anti-apoptotic gene	F67I	DAMAGING	YES
NTRK3 (neurotrophic tyrosine kinase receptor type 3)	MAPK pathway activator involved in several types of cancer and the only gene found mutated in two independent hiPSC lines	L585R	DAMAGING	YES
POLR1C (polymerase (RNA) I polypeptide C)	Sub-unit of the DNA polymerase ε complex involved in DNA replication and DNA-damage repair	P238R	DAMAGING	NO
GSG1 (Germ cell-specific gene 1 protein)	May cause the redistribution of PAPOLB from the cytosol to the endoplasmic reticulum	V252I	DAMAGING	NO
CCKBR (cholecystokinin B receptor)	This gene encodes a G-protein coupled receptor for gastrin and cholecystokinin (CCK), regulatory peptides of the brain and gastrointestinal tract.	Y238C	DAMAGING	NO
SAMD3 (Sterile alpha motif domain containing 3)		M106L	DAMAGING	YES
UBA2 (ubiquitin-like modifier activating enzyme 2)	Posttranslational modification of proteins by the addition of the small protein SUMO	E560K	TOLERATED	
TRAF6 (TNF receptor-associated factor 6)	This protein mediates the signaling not only from the members of the TNF receptor superfamily, but also from the members of the Toll/IL-1 family.	E225K	DAMAGING	NO
MARCKSL1 (Macrophage myristoylated alanine-rich C kinase substrate-like 1)	May be involved in coupling the protein kinase C and calmodulin signal transduction systems	S116F	DAMAGING	NO
CD1B (Sterile alpha motif domain containing 3)	This gene encodes a member of the CD1 family of transmembrane glycoproteins, which are structurally related to the major histocompatibility complex (MHC) proteins and form heterodimers with beta-2-microglobulin.	D51E	DAMAGING	NO
NRP1 (Neuropilin 1)	NRP1 is a membrane-bound coreceptor to a tyrosine kinase receptor for both vascular endothelial growth factor (VEGF; MIM 192240) and semaphorin (see SEMA3A; MIM 603961) family members.	G497D	DAMAGING	NO
NEK11 (NIMA (never in mitosis gene a)- related kinase 11)	The encoded protein localizes to the nucleoli, and may function with NEK2A in the S-phase checkpoint. The encoded protein appears to play roles in DNA replication and response to genotoxic stress.	R358M	TOLERATED	YES
CTSL1 (cathepsin L1)	The protein encoded by this gene is a lysosomal cysteine proteinase that plays a major role in intracellular protein catabolism.	V239I	DAMAGING	NO
ASB3 (ankyrin repeat and SOCS box containing 3)	The protein encoded by this gene is a member of the ankyrin repeat and SOCS box-containing (ASB) family of proteins. They contain ankyrin repeat sequence and SOCS box domain. The SOCS box serves to couple suppressor of cytokine signalling (SOCS) proteins and their binding partners with the elongin B and C complex, possibly targeting them for degradation.	G79S	TOLERATED	YES
ZNF16 (zinc finger protein 16)	The protein encoded by this gene contains a C2H2 type of zinc finger, and thus may function as a transcription factor.	Z59I	DAMAGING	NO

Chapter 5: The Origin of Somatic Mutations in Induced Pluripotent Stem Cells

5.1 Abstract

Recent studies have identified somatic point mutations in both coding and non-coding genomic regions of human-induced pluripotent stem cells (iPSCs)[36, 96, 97]. However, the origin of and mechanism behind these point mutations remains unclear. Here we characterize the mutational load of three human induced pluripotent stem cell lines at three separate time points derived from a single unique fibroblast cell line with an Oct4-GFP fluorescent reporter. This unique reporter allowed us to demonstrate that three distinct mutational categories exist: pre-existing rare progenitor mutations fixed due to clonal selection, pre-culture mutations developing very early during the reprogramming and expansion process, and culture mutations that occurred during iPSC expansion and became fixed. We show that these three mutation groups have distinct properties, demonstrating that mutational load in iPSCs likely arises via separate mechanisms.

5.2 Introduction

The ability to induce pluripotency in human adult somatic cells by defined transcription factor expression is a revolutionary prospect in regenerative medicine[80]. This discovery has the potential to both open new research avenues for diseases in tissue types that are difficult to obtain and to revolutionize medicine through the use

of patient-derived replacement tissue. However, due to several recent findings, concerns exist as to whether or not iPSCs are safe for clinical use.

While early studies determined that iPSCs were very similar to embryonic stem cells (ESCs), a series of recent reports have identified multiple issues potentially preventing the clinical usage of iPSCs. It has recently been shown that iPSCs contain large-scale genomic rearrangements[45], aberrant DNA methylation patterns[26], and point mutations genome-wide[36]. However, both genomic rearrangements and aberrant methylation patterns can be somewhat mitigated through use of downstream culture and selection[83, 84]. Point mutations, on the other hand, remain fixed in every single iPS cell, and exist regardless of the age of the donor, time in culture, reprogramming method, or somatic cell type used for derivation. While it is known that some mutations in iPSCs are fixed due to clonal selection of rare progenitor mutations[36, 97], the origin of and mutational process behind the remaining mutations remains unclear, as the number of mutations observed in iPSCs greatly outpaces the number expected simply from culture and clonal selection[36, 98, 99].

Here we sought to address this issue and fully characterize when iPSCs acquire mutations. To identify the exact time points at which iPSCs acquire point mutations, we derived three iPSC lines (line B, line D, and line F) from a fibroblast cell line containing an OCT4-GFP fluorescent reporter, and performed whole-genome sequencing on each at two to three separate time points (including at “passage zero,” when only 1000 pluripotent cells were present). This allowed us to divide reprogramming-associated mutations into three categories: pre-existing mutations, which were present at low levels in the progenitor population and were fixed during reprogramming; pre-culture mutations, which occurred very early during the reprogramming process; and culture mutations, which occurred during iPSC

passaging and became fixed in the population. We demonstrate that these three separate groups of mutations have unique properties, indicating that iPSC lines gain mutations via separate mechanisms.

5.3 Methods

5.3.1 hiPSC derivation

For the formation of hiPS cells OCT4-GFP+ H1-derived fibroblasts were infected with equal proportions of retroviruses encoding for OCT4, SOX2, KLF4 and c-MYC by spinfection of the cells at 1,850 r.p.m. for 1 h at room temperature in the presence of polybrene (4 μg/ml). After two serial infections, cells were passaged onto fresh MEFs and switched to hES cell medium (DMEM/F12 (Invitrogen) supplemented with 20% Knockout serum replacement (Invitrogen), 1 mM l-glutamine, 0.1 mM non-essential amino acids, 55 mM β-mercaptoethanol and 10 ng/ml bFGF (Joint Protein Central)) four days after the first infection. For the derivation of hiPS cell lines, colonies were manually picked and maintained on fresh MEF feeder layers for five passages before the growth in Matrigel/mTesR1 (Stem Cell Technologies) conditions. DNA was extracted after 30 passages for line B and after 5 and 15 passages for lines D and F.

5.3.2 Shotgun sequencing library construction (early passage only)

During reprogramming, once colonies with pre-pluripotent morphology reached a size of approximately 1000 cells and began to fluoresce, dissection and cell extraction was performed. For each green colony, approximately 700 cells were removed via manual dissection. These cells were removed, separated into three

groups, and frozen. Each colony had 300 cells left behind to continue growing into developed iPSC lines.

For those lines that developed, each stored passage zero cell set was lysed using heat/proteinase lysis, adding 5 U Protease (Qiagen) in 1x lysis buffer and incubating at 37 ºC for 10 min, 50 ºC for 15 min, and 70 ºC for 20 min. The resultant DNA from each set was divided equally into 24 or 36 separate PCR tubes. In each tube, a low-volume multiple displacement amplification reaction was performed, adding RepliPhi MDA MasterMix (Epicentre) up to a volume of 5 uL and incubating at 30 ºC for 2 hours. These amplicons were each processed into an individual shotgun sequencing library in order to ensure high total coverage across the genome at passage 0.

5.3.3 Shotgun sequencing library construction (all passages)

Shotgun sequencing libraries were generated from both the individual passage 0 amplicons and from the isolated passage 5, 15, and 30 genomic DNA using a modified version of the Nextera transposase protocol[6] (Illumina, San Diego, CA). Nextera transposase enzyme was diluted 50 fold in 1x TE buffer and glycerol. Transposase reactions were carried out in 5 uL reaction volumes, with 1 uL 5x HMW tagmentation buffer, 1 uL diluted enzyme, and 3 uL of amplicon or genomic DNA. Reactions were incubated for 5 minutes at 55 degrees C. 5 U Exo minus Klenow (Epicentre, Madison, WI) and 1 mM dNTPs were added and incubated at 37 ºC for 14 minutes followed by 65 ºC for 20 minutes. The resultant transposed products were then amplified and barcoded in a two-stage PCR reaction using 1x KAPA SYBR master mix (KAPA Biosystems, Woburn, MA), 10 uM Nextera Adapter 1, and 10 uM barcoded Nextera Adapter 2 for the first step and 1x KAPA SYBR master mix, 10 uM Illumina Primer 1, and 10 uM Illumina primer 2 for the second step. The sequencing

libraries were then purified using Ampure XP beads. As each sample contained its own individual barcode, the libraries were pooled and sequenced on an Illumina GA IIx and Illumina HiSeq.

5.3.4 Consensus sequence generation and variant calling

Variant calling was performed as previously described[36]. Briefly, reads obtained from the Illumina Genome Analyzer were post-processed and quality filtered using GERALD, mapped to the genome using BWA, downsampled using Picard and used to generate a consensus sequence for each sample using GATK. The consensus sequences were then compared to find candidate novel mutations in hiPSCs. Sites where each hiPSC line showed heterozygous SNPs not observed in the progenitor line were considered as candidate mutations if no allelic content was present in the somatic progenitor and if the candidate mutation had not previously been observed in other samples or the dbSNP database.

5.3.5 Sanger validation of candidate mutations

Out of the mutation candidates, 96 were validated using Sanger Sequencing. Genomic DNA of both the hiPSC line and its somatic progenitor (6 ng each) was amplified in separate 50 ml PCR reactions with 100 nM of specifically designed forward and reverse primers around the mutation site and 25 ml of Taq 2x master mix (NEB) at 94ºC for 2min, followed by 35 cycles of 94ºC for 30s, 57ºC for 30s and 72ºC for 30 s, and final extension at 72ºC for 3 min. The PCR products were then purified with Qiagen Qiaquick columns, and 10 ng of purified DNA was pre-mixed with 25 pmol of the forward primer for Sanger sequencing at Genewiz Inc. It was determined that mutations candidates containing at least a 30% presence of the minor allele appeared to be true positives 99% of the time.

5.3.6 Mutation characterization

After mutations were called and filtered, each was placed into one of three mutation groups based on its allele frequency in each line: pre-existing, pre-culture, and culture mutations. Pre-existing mutations were defined as those mutations containing minor allele presence in more than one iPSC line; thus, a mutation called in line B but showing minor allele presence in line F at passage 0 or line D at passage 5 would be called pre-existing. Pre-culture mutations were defined as those containing allele presence in only one iPSC line but at both passage 0 and later passages; a mutation identified only in iPSC line B at both passage 0 and at passage 30 but not found in line D or F would be called pre-culture. Culture mutations were defined as those containing allele presence in only one iPSC line but only at later passages; a mutation called in line F at passage 15 but showing no allelic presence in any other library would be called a culture mutation.

5.3.7 Association of mutation groups with epigenetic markers

Epigenetic marker presence across the genome was obtained as ChIP-Seq data from the ENCODE project. The association of each epigenetic marker to each mutation group was computed using CEAS SitePro[100], which computes the average enrichment for a given epigenetic marker across a given set of defined genomic regions (in this case mutation groups). A span of 1000 base pairs was used for each ChIP-Seq data set, with a profiling resolution of 50 bp; enrichment between mutations and epigenetic marks was examined 5 kb upstream and downstream of each site. Each enrichment profile was manually examined, and those that showed significant enrichment were extracted.

5.4 Results

5.4.1 Reprogramming-associated mutations contain three distinct categories

In order to gain a unique insight into the mutational profile of induced pluripotent stem cells, we performed reprogramming on a fibroblast line containing an OCT4-GFP fluorescent reporter. This fibroblast line, while normally not fluorescent, would fluoresce once OCT4 was expressed, allowing early detection of any cells that were beginning to acquire a pluripotent state. Early during the reprogramming experiment, we identified those pre-iPSC colonies that fluoresced. Once these colonies gained pre-pluripotent morphology and reached approximately 1000 cells in size, they were manually dissected. (Supplementary Fig. 5.1) Approximately three quarters of each colony was removed, frozen, and stored, leaving approximately 300 cells behind to continue growing. Those colonies that grew to reach a full iPSC line were then matched to their extracted “passage 0” cells; for each line, cells were extracted at either one (passage 30) or two (passages 5 and 15) downstream passages. Whole-genome sequencing libraries were then generated for each passage (Fig. 5.1). For the ~700 passage zero cells, a unique method relying on whole-genome amplification was utilized to generate an accurate library. The input DNA was subdivided into 30 separate volumes, and a modified low-bias multiple displacement amplification method was used to perform amplification; the amplicon was then converted into a sequencing library using a modified Nextera library construction protocol, resulting in a drastically reduced level of coverage dropout.

To take advantage of the unique availability of genomic DNA from a small number of input cells, we performed extremely deep whole-genome sequencing on each iPSC line at each time point and the progenitor fibroblast line, and compared the

consensus sequence of each late-passage iPSC line to that of the fibroblast progenitor. We identified an average of 790 reprogramming-associated mutations across the entire genome of each iPSC line (Table 5.1); an average of 12 mutations per iPSC line were present in coding regions of the genome, which is consistent with previously reported levels of mutational load[36, 96, 97].

Based on the presence or absence of the mutated alleles in each of the earlier iPSC passages, for the first time, we were able to divide the identified mutations into three categories: pre-existing mutations, which are present at low levels in progenitor cells and fixed during reprogramming; pre-culture mutations, which appear to be fixed in the iPSC line at passage zero but do not appear to be clonally selected from progenitor cells; and culture mutations, which arose during later stage iPSC growth and expansion. Pre-existing and pre-culture mutations were both identified by their presence at passage zero; pre-existing mutations were also identified in multiple iPSC lines rather than being limited to one line. On average, 57% of mutations in each iPSC line were pre-existing, 26% were pre-culture, and 16% were culture mutations (Table 5.1). These three categories represent unique categories of reprogramming-associated mutations that seem to have occurred at different time points during the reprogramming process. However, it is possible that some mutations categorized as pre-culture might in reality be pre-existing mutations that were present at such low levels that only one iPSC line inherited them. We performed further analysis to elucidate whether these three mutation groups were truly distinct.

5.4.2 Mutation categories contain unique properties

In order to determine whether these three groups of reprogramming-associated mutations represented unique mutational processes, we analyzed the

specific base pair changes that occurred in each mutational group. It has been previously demonstrated that mutational processes tend to introduce different types of base pair changes; for example, ultraviolet light commonly introduces C to T mutations[101], while oxidative damage commonly introduces G to T mutations[102]. We posited that if similar mutational processes were at work for each category, a similar base change profile would be observed for each group. We observed that while pre-existing and culture mutations had a similar mutational profile, pre-culture mutations had a completely separate mutational profile dominated by C to A / G to T transversions (Figure 5.2). This indicates that while pre-existing and culture mutations appear to arise by the same process (likely mutations due to cellular expansion *in vitro*), pre-culture mutations seem to arise from a separate source.

Previous studies have demonstrated that in several disease cases, germline and somatic mutations both tend to be enriched in genomic regions with distinct epigenetic signatures[103, 104]. To determine if the three types of reprogramming-associated mutations possessed enrichment in separate epigenetic categories, we performed an enrichment analysis between each mutation group and histone marker data from ENCODE[105] using SitePro[100]. While no strong enrichment was observed between mutations and any histone marks, we determined that pre-existing mutations tended to occur in regions with high DNAse I accessibility in Fibroblasts and hESCs; on the other hand, pre-culture and culture mutations did not have any strong enrichment with respect to accessibility. (Figure 5.3) Based on this result, it appears that pre-existing mutations, pre-culture mutations, and culture mutations are occurring under separate conditions or through separate mutational processes, as they tend to localize separately.

5.5 Conclusions

Based on these findings, it appears that each mutation group could be a distinct category. Pre-existing and culture mutations tend to occur at adenine or thymine bases, while pre-culture mutations tend to occur at cytosine or guanine bases; both pre-existing and culture mutations are also dominated by transitions rather than transversions. It is therefore possible that these mutations might be occurring through similar mechanisms, with pre-existing mutations occurring before reprogramming and culture mutations occurring afterward. Further study of DNAse I in iPSC lines could reveal whether or not culture mutations are similarly localizing in accessible genomic regions.

However, despite the evidence that the observed mutational categories are distinct, a large amount of variability still exists within each mutation group. Thus, despite the observed trends, it is not possible with current data to separate mutations from an iPSC line into each category without performing sequencing at passage zero. It is additionally possible that mutations classified as “pre-culture” might in fact be pre-existing mutations that were present at extremely low levels in the progenitor cells. However, the disparate base change profiles seem to dismiss this possibility and imply separate mutational mechanisms. Sequencing of additional iPSC lines at passage zero might allow further characterization of these trends.

Taken together, these results indicate that reprogramming-associated point mutations are accumulated genome-wide by induced pluripotent stem cells throughout progenitor culture, reprogramming, and iPSC culture. Mutations occur at a high rate, with an average of one reprogramming-associated mutation approximately every four million base pairs; this rate is on the same order of magnitude as that observed in cancer lines[106] (which has approximately one point

mutation per one million base pairs). Mutations appear to occur in three separate categories, and a majority of mutations appear to be inherited from rare progenitor mutations; this is consistent with previously reported results of iPSC whole genome sequencing[96, 97]. Thus, only pre-culture mutations, comprising a mere 26% of mutations, have the possibility of being eliminated by improvements to the reprogramming process. These results pose potential problems for the clinical use of induced pluripotent stem cells. While previous results have indicated that most mutations do not appear functional for reprogramming, it is possible that iPSC-derived cells might behave erratically due to mutations having a functional effect in other tissue contexts. In order for iPSCs to become clinically relevant, a functional test guaranteeing a specific iPSC line's safety in a given tissue context must be developed.

5.6 Acknowledgements

We thank Athanasia Panopoulos for performing reprogramming and cell culture and Rui Liu for performing library construction experiments. A. Gore is supported by the Focht-Powell Fellowship and a CIRM predoctoral fellowship. Work in this chapter was supported by NIH R01 HL094963 and a UCSD new faculty start-up fund to Kun Zhang, and grants from Fundacion Cellex, TERCEL-ISCIII-MINECO, Sanofi, National Institutes of Health and the G. Harold and Leila Y. Mathers Charitable Foundation to Juan Carlos Izpisua Belmonte.

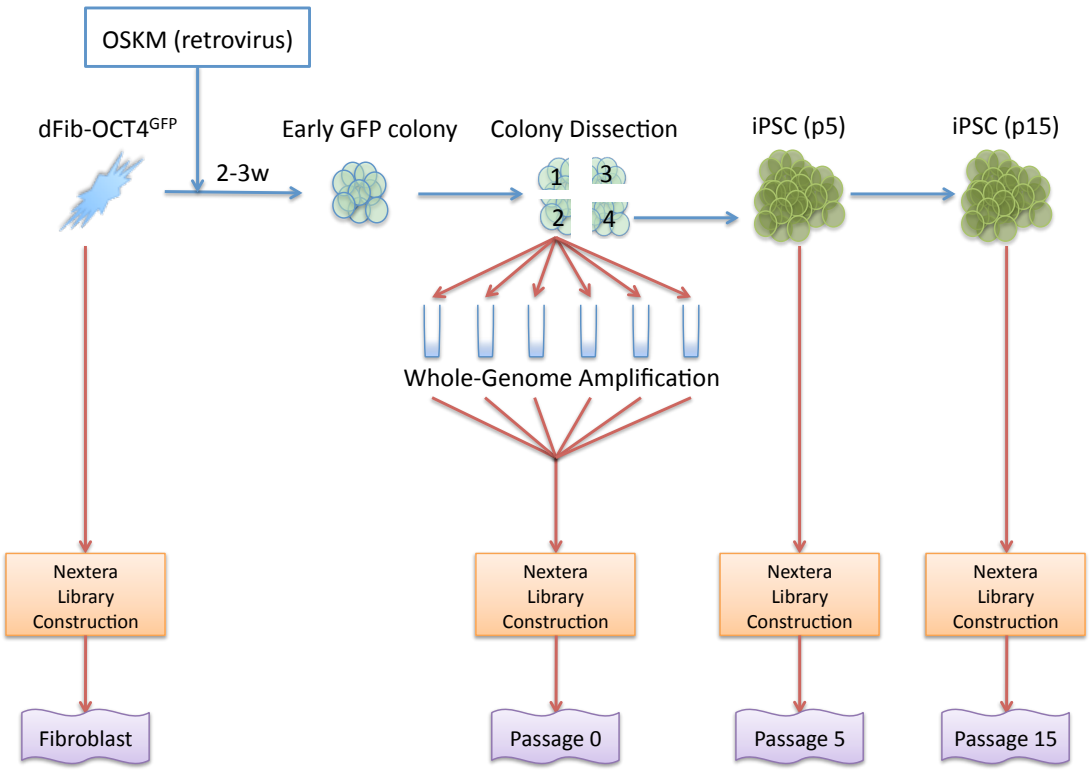


Figure 5.1. Schematic of experimental workflow allowing characterization of the origin of reprogramming-associated mutations. Fibroblasts containing an OCT4^{GFP} reporter were reprogrammed using the traditional OSKM retroviral method. During the reprogramming experiment, once colonies demonstrating pluripotent-like morphology and expressing GFP had reached approximately 1000 cells in size, three quarters of each colony was removed. The remaining quarter was grown and passaged into an adult iPSC line. After the adult iPSC line was established, stored cells from each time point were matched and DNA was extracted from each set. Passage zero DNA was subdivided into 24-36 individual reactions and was whole-genome amplified by Multiple Displacement Amplification. Nextera library construction was then utilized on amplicons from passage 0 iPSCs and genomic DNA from fibroblasts and adult iPSCs to generate Illumina-compatible sequencing libraries.

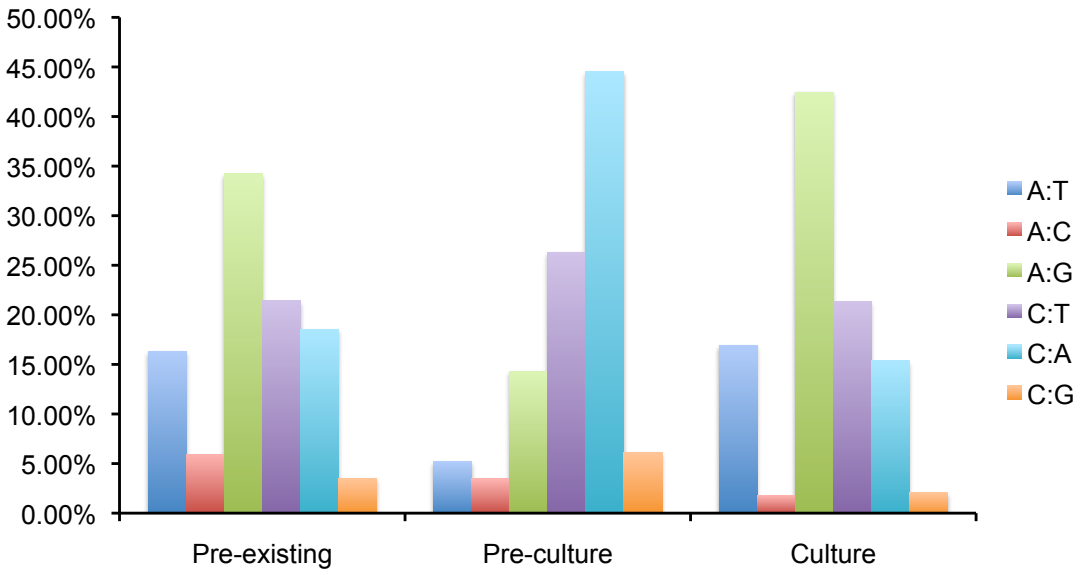


Figure 5.2. Distribution of base changes observed in pre-existing, pre-culture, and culture mutations.

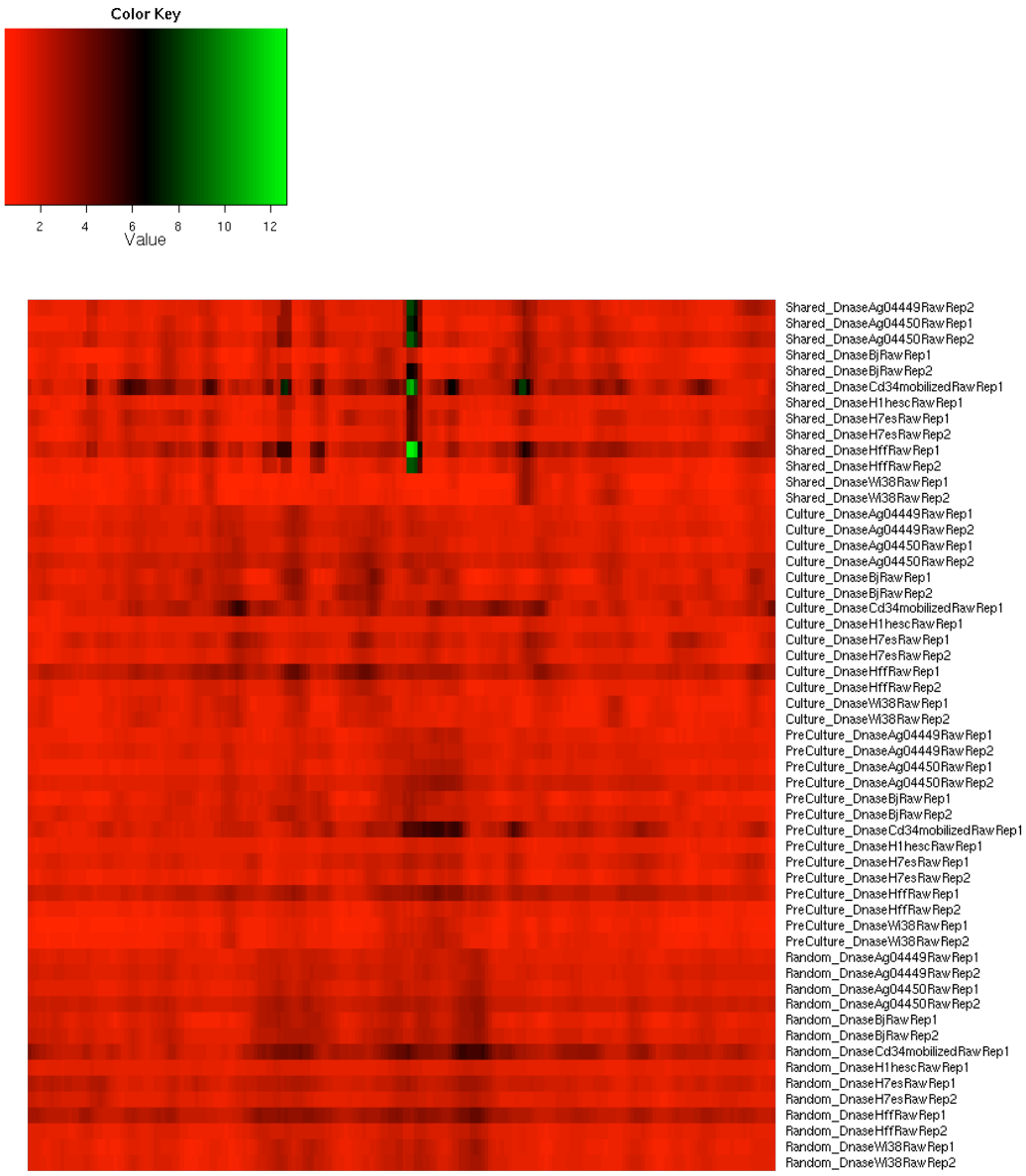
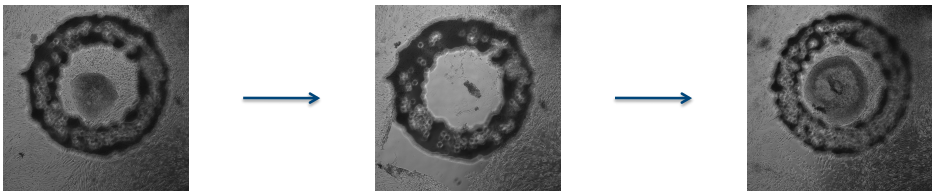


Figure 5.3. Association between DNAse I accessibility and mutation sites. The horizontal axis represents 5 kb upstream and downstream of a mutation site. Green represents accessibility of the genome to DNAse I, while red represents inaccessibility. Pre-existing mutations (top section) tend to occur in DNAse I accessible areas from human embryonic stem cells and fibroblasts. No strong trends are present in terms of upstream or downstream accessibility to DNA.



Supplementary Figure 5.1. Colony dissection of a growing iPSC line. Growing iPSC colony (outlined with black marker) showing GFP expression and pluripotent morphology was allowed to reach a size of ~1000 cells. Approximately 3/4 of the colony was removed by manual dissection; extracted cells were spun down into a pellet and then frozen for storage. A few hundred cells were left behind. In just a few days, the colony regrew to its former size; this colony continued to expand and was eventually passaged into an iPSC line.

Table 5.1. Mutations in iPSC lines. Whole-genome sequencing identified a similar number of mutations in each of the three iPSC lines.

Cell Line	Passage	Genomic Coverage	Number of Mutations			
			Pre-existing	Pre-culture	Culture	TOTAL
Fibroblast		38x				
iPSC Line B	0	85x	453	227	99	779
	30	31x				
iPSC Line D	0	54x	453	210	158	821
	5	22x				
	15	44x				
iPSC Line F	0	57x	453	194	124	771
	5	17x				
	15	38x				
AVERAGE		43x	453	210	127	790

Chapter 6: Identification Of a Specific Reprogramming-Associated Epigenetic Signature in Induced Pluripotent Stem Cells

6.1 Abstract

Generation of human induced pluripotent stem cells (hiPSCs) by the expression of specific transcription factors depends on successful epigenetic reprogramming to a pluripotent state. Although hiPSCs and human embryonic stem cells (hESCs) display a similar epigenome, recent reports demonstrated the persistence of specific epigenetic marks from the somatic cell type of origin and aberrant methylation patterns in hiPSCs. However, it remains unknown whether the use of different somatic cell sources, encompassing variable levels of selection pressure during reprogramming, influences the level of epigenetic aberrations in hiPSCs. In this work, we characterized the epigenomic integrity of 17 hiPSC lines derived from six different cell types with varied reprogramming efficiencies. We demonstrate that epigenetic aberrations are a general feature of the hiPSC state and are independent of the somatic cell source. Interestingly, we observe that the reprogramming efficiency of somatic cell lines inversely correlates with the amount of methylation change needed to acquire pluripotency. Additionally, we determine that both shared and line-specific epigenetic aberrations in hiPSCs can directly translate into changes in gene expression in both the pluripotent and differentiated states. Significantly, our analysis of different hiPSC lines from multiple cell types of origin allowed us to identify a reprogramming-specific epigenetic signature comprised of

nine aberrantly methylated genes that is able to segregate hESC and hiPSC lines regardless of the somatic cell source or differentiation state.

6.2 Introduction

Induction of pluripotency in human somatic cells is an inefficient process that can be achieved by the expression of defined transcription factors[44, 55, 80-82]. This reprogramming process involves global epigenetic remodeling and overcoming similar roadblocks present during cell transformation, which might affect genomic and epigenomic integrity[107]. In fact, several recent reports have shown that human induced pluripotent stem cells (hiPSCs) contain genetic and epigenetic aberrations throughout their genome compared with their parental somatic cell populations or to human embryonic stem cells (hESCs)[26, 36, 39, 45, 83, 84]. For example, the analysis of whole-genome DNA methylation profiles at single-nucleotide resolution in hiPSCs, their somatic cells of origin, and hESCs revealed the presence of more than 1,000 differentially methylated regions (DMRs) between hiPSCs and hESCs[26]. Moreover, this analysis, and many others, demonstrated both the persistence of specific epigenetic marks from the somatic cell of origin (residual methylation) and the acquisition of unique methylation patterns in mouse iPSCs (miPSCs) and hiPSCs[26, 85, 108-115]. Interestingly, hiPSC lines also show incomplete reprogramming of non-CG methylation in regions proximal to telomeres and centromeres[26]. Altogether, these epigenetic aberrations might explain some of the observed transcriptional variation between hESC and hiPSC lines[116-118]. In one of the most comprehensive reports to date, Bock et al.[117] characterized a panel of 20 hESC and 12 hiPSC lines to demonstrate that despite their global similarity, a number of genes in each pluripotent cell line deviated from the normal expected variation compared with the DNA

methylation and gene expression levels observed in the other pluripotent cell lines. Interestingly, they reported that no apparent epigenetic deviation was unique to all hiPSC lines[117]. Altogether, these findings demonstrate that hiPSCs contain epigenetic aberrations. However, a majority of these reports predominantly used fibroblast-derived hiPSC lines and, thus, it remains unknown whether the use of alternative somatic cell types with variable levels of selection pressure for reprogramming might result in hiPSC lines containing fewer (or perhaps none) of these epigenetic alterations. Furthermore, although it has been shown that aberrantly methylated CpG sites are transmitted to differentiated cells[26], it remains unclear whether these epigenetic aberrancies result in transcriptional variation after differentiation.

In this work, we characterize at single nucleotide resolution the methylation profile of 17 hiPSC lines derived from six different somatic cell types with varied reprogramming efficiencies. Our results show that, independent of the somatic cell source used for reprogramming, all hiPSC lines analyzed contain abnormal epigenetic patterns. We determine that a majority of these aberrantly methylated CpG sites are transmitted to differentiated cells and can be associated with changes in gene expression after differentiation. Importantly, we identify a reprogramming-associated epigenetic signature comprised of nine aberrantly methylated genes that can segregate hESC and hiPSC lines both in the pluripotent state and after differentiation. These observations will contribute to a deeper understanding of the reprogramming process and underscore the need for a rigorous evaluation of the epigenetic integrity of hiPSC lines.

6.3 Methods

6.3.1 hiPSC Generation

The hiPSC lines ASThiPS4F1, ASThiPS4F2, ASThiPS4F3, ASThiPS4F4, ASThiPS4F5, HUVhiPS4F1, HUVhiPS4F3, HUVhiPS4F6, FhiPS4F2, FhiPS4F5, FhiPS4F7, KhiPS4FA, PGP1-iPS, and NSChiPS2F have been previously described[36, 86-88, 119], and were obtained from existing cultures. To generate hiPSCs (KhiPS4F8, MSChiPS4F4, and MSChiPS4), experiments were performed as described with minor modifications[51]. Briefly, keratinocytes or MSCs were infected with an equal ratio of retroviruses by spinfection of the cells at 800x g for 1 h at room temperature in the presence of polybrene (4 μg/mL). After two (for keratinocytes) or three (for MSCs) viral infections, cells were trypsinized and transferred onto fresh irradiated mouse embryonic or human fibroblasts (iMEFs or iHFs), respectively. One day after, cells were switched to hESC medium (DMEM/F12 or KO-DMEM (Invitrogen) supplemented with 20% Knockout Serum Replacement (Invitrogen), 1 mM L-glutamine, 0.1 mM nonessential amino acids, 55 μM β-mercaptoethanol and 10 ng/mL bFGF (Joint Protein Central)). For the derivation of hiPSC lines, colonies were manually picked and maintained on fresh mouse embryonic fibroblasts (MEFs) feeder layers for five passages before growth in Matrigel/mTesR1 conditions.

6.3.2 hiPSC Details

Three of the somatic cell types (human umbilical vein endothelial cells (HUVECs), astrocytes, and neural stem cells (NSCs)) were of fetal/neonatal origin, keratinocytes were obtained from a 5-y-old individual, mesenchymal stem cells (MSCs) were obtained from liposuctioned tissue from aged women and Fibroblasts (HFFxF and PGP1F) were obtained from two different biopsies from a 55-y-old and a

5-y-old individual. Young cell sources were preferentially chosen to rule out age-accrued DNA damage or epigenetic alterations as a possible source of aberrations, because these cell samples likely possess a lower level of exposure to natural stress and mutagenic agents. Moreover, embryonic and adult stem cells (NSCs and MSCs) were specifically chosen because they have more effective mechanisms of genomic preservation than somatic cells. All hiPSC lines were generated by using retroviral or lentiviral infection to express between two and four of the reprogramming factors. All lines were fully characterized in terms of pluripotent gene expression, transgene silencing, karyotype, and in vitro and in vivo differentiation into tissues from all three embryonic germ layers[36, 86-88, 119].

6.3.3 Immunostaining

Immunofluorescence analysis for the detection of pluripotent markers in hiPSCs or for the detection of differentiation-associated markers in teratomas was performed as described[88].

6.3.4 RNA Isolation and Real-Time PCR Analysis

Total RNA was isolated by using TRIzol Reagent (Invitrogen) according to the manufacturer's recommendations. cDNA was synthesized by using the SuperScript II Reverse Transcriptase kit for RT-PCR (Invitrogen) or the RT Supermix M-MuLV kit (BioPioneer). Real-time PCR was performed by using SYBR-Green PCR Master mix (Applied Biosystems) in a ViiA 7 Real-Time PCR System (Applied Biosystems). GAPDH expression was used to normalize values of gene expression, and data are shown as fold change relative to the value of the sample control. All of the samples were done in triplicate.

6.3.5 Teratoma Formation and Karyotype Analysis

Severe combined immunodeficient mice (NOD.Cg-Prkdcscid Il2rgtm1Wjl/SzJ; Jackson Laboratories) were used to test the teratoma induction capacity of the hiPSC lines as described[88]. hiPSC lines grown on Matrigel were processed to perform karyotype analysis as described[88]. All animal experiments were conducted by following experimental protocols approved by the Institutional Ethics Committee on Experimental Animals at the Parc de Recerca Biomedica de Barcelona (PRBB), in full compliance with Spanish and European laws and regulations.

6.3.6 Bisulfite Padlock Probe Production

Oligonucleotides were synthesized by ink-jet printing on programmable microarrays (Agilent Technologies) and released to form a combined library of 330,000 oligonucleotides. The library was prepared for padlock capture by using a described protocol[11, 120].

6.3.7 Sample Preparation and Capture

Genomic DNA was extracted by using the ALLPrep DNA/RNA Mini kit (Qiagen) and QIAamp DNA Micro kit (Qiagen). The bisulfite conversion and capture reactions were carried out on 1–1.2 µg of each sample by using established protocols[11, 120]. Briefly, DNA was bisulfite converted by using the EZ-96Methylation Gold Kit (Zymo Research). Approximately 200–300 ng of converted gDNA from each sample was captured using prepared padlock probe oligonucleotides, resulting in a circular DNA library of targeted CpG sites.

6.3.8 Bisulfite Sequencing Library Construction

The circular DNA library was amplified as described[11] with slight modifications. Briefly, two-thirds of each capture reaction was used to prepare two real-time PCR

reactions with 20 pmol each of AmpF6.4Sol and AmpR6.3 indexing primers and 50 µL of 2x KAPA SYBR FAST Universal qPCR Master Mix (KAPA Biosystems). Thermocycling was carried out at 98 °C for 30 s; 8 cycles of 98 °C for 10s, 58°C for 30s, and 72°C for 30s; and 12–14 cycles of 98°C for 10 s and 72 °C for 30 s. Finally, the reactions were held at 72 °C for 3 min. Duplicate reactions were then pooled, purified using 0.8x AMPure magnetic beads (Agencourt), and quantified by 6% polyacrylamide gel electrophoresis. Samples were mixed in equimolar ratios to create two libraries, and together were size selected with 6% polyacrylamide gel electrophoresis. The first pool with 60 libraries was sequenced in five lanes of a paired-end 100 bp Illumina Hi-Seq run, and the second pool with 12 libraries was sequenced in 1 lane of a paired-end 110-bp Illumina Hi-Seq run.

6.3.9 Bisulfite Read Mapping and Data Analysis

Bisulfite converted data were processed as described[11, 38]. Heatmaps and dendrograms were created from the Pearson's correlation matrices of the relative change in methylation level between each hiPSC line and its somatic progenitor, and the absolute methylation level at each site in each line.

6.3.10 Statistical Analysis: Identification of Differentially Methylated Sites

To identify sites showing a change in methylation after reprogramming, a $\chi 2$ test with Yates' correction was carried out on each CpG site characterized in each hiPSC line and corresponding paired somatic cell line. The Benjamini–Hochberg method was used to correct for multiple testing errors; the false discovery rate (FDR) was set at 1%. This resulted in a set of differentially methylated sites (DMSs) for each hiPSC line; at each site, the methylation level was statistically significantly different from the somatic progenitor line and different by at least a 0.2 change in absolute

methylation level. A set of 5,701 DMSs were shared by all 17 hiPSC lines and were split into two groups (hypermethylated or hypomethylated) based on the mean change in relative methylation level between somatic progenitor and hiPSC line. A total of 5,056 sites were hypermethylated and 645 sites were hypomethylated in all hiPSC lines after reprogramming. Each list of sites was tested for functional similarity by using GREAT (http://great.stanford.edu), along with a list of the 336,904 sites characterized in all lines as background. The single closest gene within 10 kb of a DMS from each list, and the enriched GO Biological Process terms chart were generated by using GREAT.

6.3.11 Statistical Analysis: Identification and Classification of Epigenetic Aberrations

The following procedure was followed for all hiPSC lines to first identify aberrant CpG sites and then categorize them as residual methylation or de novo methylation. First, the methylation levels at each CpG site in the hiPSC line were considered and compared with the average methylation level and the upper and lower bounds of methylation level for the same site in the hESC lines. Those sites showing at least a 0.2 change in absolute methylation level, considered to have methylation levels from different underlying distributions by the $\chi 2$ test (with Benjamini–Hochberg multiple testing correction; FDR = 0.01), and having a methylation level at least 0.2 away from either the maximum or the minimum hESC methylation level were considered to be “aberrantly methylated.” These aberrant CpG sites were then classified into two categories: de novo methylation and residual methylation. Sites were classified as de novo methylation if the methylation level met three conditions: the level in the hiPSC line was statistically significantly different by the same $\chi 2$ criteria than the level in its corresponding somatic cell progenitor, the

hiPSC line's absolute methylation level was at least 0.2 away from the somatic cell line's, and the hiPSC methylation level was not between that of its somatic progenitor and the hESC lines. Other aberrant sites were classified as residual methylation. CpG sites were associated with a gene if the gene's transcription start or end site was located within 10 kilobases; CpG sites located more than 10 kb away from a gene were considered to be "unlinked." In cases where multiple genes were within 10 kb, the CpG site was associated with the closest gene.

6.3.12 Classification of Unique and Shared Epigenetic Aberrations

To obtain an enriched list of genes and their associated CpG sites for functional analysis, genes that showed either "shared" or "line-specific" residual methylation and de novo methylation patterns were identified. In order for a gene to have been considered to carry "shared" residual methylation or de novo methylation patterns, it must have contained CpG sites showing residual methylation or de novo methylation in at least 16 of the 17 analyzed hiPSC lines. In order for a gene to have been considered to carry "line-specific" residual methylation or de novo methylation patterns for a given hiPSC line, it must have contained CpG sites showing residual methylation or de novo methylation in no more than three other lines derived from separate progenitor cell types. This grouping resulted in lists of genes that showed aberrant methylation in either all hiPSC lines or only a few hiPSC lines and allowed us to focus on genes that could have methylation-based functional changes in expression.

6.3.13 Activin-Induced Differentiation

Cells were treated with media (mTeSR1) plus Activin-A (100 ng/mL). Media was replaced daily for 5 d. Cells were then collected with TrypLE (Invitrogen), washed with PBS, and processed for DNA and RNA isolation.

6.3.14 BMP4-Induced Differentiation

Cells were treated with BMP4 for 5 d with minor modifications as described[121]. Cells were then collected with TrypLE (Invitrogen), washed with PBS, and processed for DNA and RNA isolation.

6.3.15 Analysis of Epigenetic Aberrations after Differentiation

Targeted bisulfite sequencing was performed on genomic DNA from pluripotent and differentiated cultures of hiPSC and hESC lines as described above. Aberrant methylation was called in the pluripotent hiPSC state as described above. Two comparisons were then performed for each aberrantly methylated site by using the above method to call differential methylation: the first between the hiPSC-pluripotent state and the hiPSC-differentiated state, and the second between the hESC-differentiated state, and the hiPSC-differentiated state. Based on the results of these two statistical tests, each site was characterized into one of four categories: (i) hiPSC-pluripotent and hiPSC-differentiated states are similar, but different from hESC-differentiated state; (ii) hiPSC-pluripotent state, hiPSC-differentiated, and hESC-differentiated states are all similar; (iii) hiPSC-differentiated and hESC-differentiated states are similar, and different from hiPSC-pluripotent state; or (iv) all three states are different.

6.3.16 Microarray Data

The GeneChip microarray (Affymetrix ST 1.0 microarrays) processing was performed according to the manufacturer's protocols (Affymetrix). The amplification and labeling were processed as indicated in Nugen protocol with 100 ng of starting RNA. For each sample, 3.75 mg of ssDNA were labeled and hybridized to the Affymetrix ST 1.0 chips. Expression signals were scanned on an Affymetrix GeneChip Scanner. The data extraction was done using the Affymetrix GCOS software. The data analysis was performed by using the affylmGUI package in R-Bioconductor. Briefly, .CEL files were imported in R-Bioconductor for preprocessing and normalization. Cluster 3.0 software was used to perform hierarchical clustering on RMA-normalized probeset intensity values. The array as well as the methylation data reported in this paper have been deposited in the Gene Expression Omnibus database under accession numbers GSE39210 and GSE40372.

6.4 Results

6.4.1 Reprogramming Efficiency Inversely Correlates with the Percentage of Epigenetic Modifications Observed After Reprogramming

To gain insight into the epigenetic integrity of hiPSCs, we performed targeted bisulfite sequencing with padlock probes[11, 120] to analyze the methylomes of 17 hiPSC lines, their 6 somatic cell types of origin, and 7 hESC lines (Methods, Table 6.1). We designed and synthesized a set of 330,000 synthetic probes targeting ~140,000 genomic regions known to be differentially methylated across different cell types[39-41] and additional functional regions. We determined the absolute methylation levels for an average of ~529,000 CpG sites per sample (Table 6.1). Although only ~1% of the human genome was covered by this assay, these preselected CpG sites were more

than twice as informative as typical sites in CpG islands characterized by using lower resolution sequencing or in previously used bisulfite sequencing methods (Supplementary Table 6.1). Unbiased hierarchical clustering of global methylation levels demonstrated a clear segregation of somatic cells and pluripotent cells (Fig. 6.1a). We also observed that hiPSC lines originating from the same somatic cell type tended to cluster together in subgroups (Fig. 6.1a,b), which, as reported[26, 85, 108-115], supports the existence of residual methylation from somatic cells of origin in hiPSCs.

We analyzed the number of differentially methylated CpG sites (DMSs) in each hiPSC line by comparing each cell line to its direct somatic cell source of origin (Supplementary Table 6.1). We observed that between 23% and 37% of CpG sites analyzed underwent a change in methylation state, with mesenchymal stem cells (MSCs) and fibroblasts requiring the most dramatic epigenetic change following reprogramming and neural stem cells (NSCs) requiring the least (Supplementary Table 6.1). Interestingly, the percentage of DMSs after reprogramming correlated inversely with reprogramming efficiency, with cell sources undergoing the fewest epigenetic modifications reprogramming at higher efficiency (Fig. 6.1c). Moreover, we confirmed previous findings[26] and determined that, independent of somatic cell source, the global change in methylation observed after reprogramming is toward a more methylated state (Supplementary Fig. 6.1a). Next, we investigated whether different somatic cell sources shared a core set of DMSs that might be essential to epigenetically reprogram to a pluripotent state. In fact, we observed that ~5,700 DMSs were shared among all hiPSC lines (Supplementary Fig. 6.1b). Analysis of Gene Ontology for genes that could potentially be regulated by these DMSs revealed that genes with hypomethylated DMSs appeared to be enriched for cell signaling, protein refolding, cell metabolism, and neuronal development, whereas genes with

hypermethylated DMSs appeared to be enriched for cell-cell adhesion and receptor behavior (see online version[122]).

6.4.2 hiPSC Lines Share a Core Set of Aberrantly Methylated Genes that Segregate them from hESCs

We compared the methylation state at each CpG site in individual hiPSC lines to that of their parental source and seven hESC lines. Using an algorithm based on the χ2 test with multiple testing corrections, we identified sites where hiPSC lines carried a methylation pattern significantly different from hESC lines (Methods). hiPSC lines derived from the same somatic cell source carried similar, although not identical, aberrant methylation patterns and clustered together based on methylation level at aberrant sites (Fig. 6.2a). We categorized the aberrantly methylated CpG sites into two categories: residual methylation, where the CpG site in a hiPSC line retains the methylation level of its parental cell instead of reaching the level observed in hESCs (Fig. 6.2b), and de novo methylation, where the CpG site in a hiPSC line acquires a methylation state found neither in its somatic source nor in hESCs (Fig. 6.2b). We determined that the percentage of aberrant CpG sites varied between 0.92% and 3.82% across the hiPSC lines analyzed. Furthermore, the percentage of CpG sites that showed residual methylation or de novo methylation varied between 0.32% and 1.60% and 0.57% and 2.98%, respectively (Table 6.1). Although we did not find a direct correlation between the amount of aberrant methylation and reprogramming efficiency or somatic cell type, we noted that some cell types appeared to possess lower aberrant methylation levels (e.g., astrocyte-derived lines) compared with others (e.g., fibroblast-derived lines) (Table 6.1). We determined that most aberrantly methylated CpG sites showing de novo methylation were characterized by excessive methylation after reprogramming (Fig. 6.2c), whereas most aberrantly methylated

CpG sites associated with genes showing residual methylation were characterized by only partial methylation occurring after reprogramming (Fig. 6.2d).

To gain insight into potential functional consequences of these epigenetic aberrations, we linked each aberrant CpG site with its closest gene (Methods) and used this subset of genes for further analysis. Interestingly, we observed that a very small number of genes contained aberrant methylation patterns in nearly all hiPSC lines assayed in our study (16/17 hiPSC lines) regardless of somatic cell source (see online version[122]). We hypothesized that the nine genes (PTPRT, TMEM132C, TMEM132D, TCERG1L, DPP6, FAM19A5, RBFOX1, CSMD1, and C22ORF34) we identified might represent a core set of aberrantly methylated genes that can systematically distinguish hiPSC and hESC lines. Thus, we performed unbiased hierarchical clustering based on the methylation status of CpG sites associated to this small subset of genes in previously published independent methylation datasets. Specifically, we first examined a set of whole-genome bisulfite sequencing data performed in three hESC and five hiPSC lines[26]. We found that, similar to what we observed for our dataset, the methylation level of CpG sites associated to the nine genes was able to clearly segregate hESC and hiPSC lines into two distinct groups (Fig. 6.2e). Additionally, we used a recently published dataset that profiled the genome-wide DNA methylation level for more than 450,000 CpG sites in 19 hESC and 29 hiPSC lines[108] and observed that, despite the lower resolution, a similar clustering analysis clearly segregated all but two hiPSC lines from hESC lines (Fig. 6.3a).

Next, we investigated whether our core set of aberrantly methylated genes showed differential gene expression in hiPSC lines compared with hESC lines by performing real-time PCR analysis on RNA obtained from six hiPSC lines and six

hESC lines. An unbiased hierarchical clustering of the real-time PCR data results examining the gene expression of the nine shared aberrantly methylated genes demonstrated a clear segregation between hiPSC and hESC lines (Fig. 6.2f). Furthermore, to determine the global relevance of these findings, we also performed a similar unbiased hierarchical clustering by using previously reported independent datasets containing a variety of hESC and hiPSC lines (a total of 12 datasets). Overall, when examining the expression of these nine genes, we determined that although clear outliers and different subgroups among hiPSC lines were detected, a majority of the dataset clusters showed separation between hiPSC and hESC lines (see online version[122]). These combined results suggest the existence of shared epigenetic aberrancies associated to a small subset of genes in hiPSC lines. The validation of these aberrancies by using our data and data from independent laboratories strongly corroborates the strength of our findings.

6.4.3 Aberrant Methylation at CpG Sites is Transmitted During hiPSC Differentiation, Resulting in Transcriptional Changes Compared with Differentiated hESCs

To further test whether the aberrant methylation and gene expression levels observed in hiPSC lines were maintained after loss of the pluripotent state, we differentiated five hESC lines and five hiPSC lines toward two different germ cell layers, endoderm and trophoectoderm, by using Activin-A and BMP4, respectively. We then performed targeted bisulfite sequencing to analyze the methylomes of the hESC and hiPSC lines in their pluripotent and differentiated states. In addition, the gene expression levels of H9, HUVhiPS4F1 and HUVhiPS4F3 were profiled in duplicate by using Affymetrix ST 1.0 microarrays. Between 0.3% and 1% of CpG sites were aberrantly methylated in the hiPSC lines with respect to hESC lines (see online

version[122]). We first investigated whether these epigenetic aberrations resulted in changes in gene expression in undifferentiated cells. We observed that between 3% and 7% of genes linked to these aberrantly methylated sites showed differential gene expression in hiPSC lines compared with hESC lines (see online version[122]). Additionally, we tested the expression of five genes with line-specific epigenetic de novo methylation in HUVhiPS4F1 and observed that these genes also showed differential gene expression compared with other hiPSC or hESC lines (see online version[122]). Taken together, these results indicate that some epigenetic aberrations are associated with changes in gene expression levels.

We next analyzed the methylation status of the aberrant CpG sites in both hiPSCs and hESCs after each differentiation protocol. The CpG sites were classified based on their post-differentiation methylation status into four categories (Fig. 6.3, see online version[122]). We observed that ~20–50% of the aberrantly methylated CpG sites detected in hiPSC lines remained aberrant after differentiation into either of the two separate cell lineages (Fig. 6.3a). Importantly, we observed that a subset of genes associated with these CpG sites showing differential gene expression level in undifferentiated hiPSCs compared with hESCs still remain in that condition regardless of differentiation protocol (Fig. 6.3b and online version[122]). Finally, to further validate the potential of the identified hiPSC-specific epigenetic signature described above, we clustered the pluripotent cells and their differentiated progenies based on both the methylation level and transcriptional abundance of the nine signature genes (Fig. 6.3c,d and online version[122]). Interestingly, the samples segregated based on whether the progenitor line was a hiPSC or hESC, and clustered by specific cell line but not by differentiation protocol. Altogether, these data suggest that the methylation and gene expression levels of the aberrantly methylated

genes in hiPSC lines still segregate hESCs and hiPSCs even after differentiation toward independent germ cell layers.

6.5 Conclusions

In this work, we have used an expanded bisulfite padlock probe set to interrogate the methylation level of targeted CpG sites identified to carry differential methylation in various cell states regardless of CpG density[39-41, 120]. This unique approach identified genes linked to individual aberrantly methylated CpG sites that are not necessarily located in CpG-enriched genomic regions. Our results show that epigenetic aberrations occur in hiPSCs regardless of the somatic cell type of origin. We demonstrated that aberrant epigenetic patterns in hiPSC lines influence gene expression and could explain functional diversity within hiPSC lines and between hiPSC and hESC lines[93-95]. In fact, we observed the existence of genes aberrantly methylated and differentially expressed in hiPSC lines compared with hESCs that still remained in that condition after differentiation regardless of differentiation protocol.

The use of hiPSC lines derived from six different somatic cell types enabled us to narrow down a precise core set of genes that contained aberrant epigenetic patterns associated with the hiPSC state. This analysis led us to identify a reprogramming-associated epigenetic signature based on the methylation level of nine genes that could segregate hESC and hiPSC lines in both the pluripotent state and after differentiation. There have been many reports suggesting the existence of epigenetic and transcriptional differences between hiPSC and hESC lines[26, 39, 85, 108-118]. Interestingly, recently reported analysis using restricted representation bisulfite sequencing (RRBS) showed that although cell line-specific outliers at both the methylation and gene expression levels could be identified, no apparent epigenetic

deviation was unique to all hiPSC lines[117]. However, the data presented therein[117] did not appear to target any of the aberrantly methylated CpG sites covered in our hiPSC-specific signature, because RRBS mainly focuses on the analysis of CpG islands (resulting in low coverage of genomic regions with low CpG density, including many functional elements such as enhancers). When we compared the lists of CpG sites associated with the nine genes characterized by our dataset to the Bock et al. dataset[117], there was almost no overlap between the two sets of analyzed CpG sites. In fact, in the Bock et al. dataset[117], only 1 CpG site of the ~600 we identified as aberrantly methylated CpG sites associated to the 9 genes was included in their analysis. Thus, when we clustered the pluripotent cell lines used in the Lister et al. dataset[26] (which analyzed a near-complete selection of CpG sites genome-wide in an unbiased manner) based on the CpG sites that were analyzed by the Bock et al. dataset[117], no clear separation was observed between hESC and hiPSC lines. However, when we clustered the hESC and hiPSC lines included in the Lister et al. dataset based on the CpG sites analyzed in our study, we found that we were able to segregate the two different pluripotent cell types. Furthermore, when we compared our data to an extensive set of genome-wide DNA methylation profiling of hESC and hiPSC lines that had analyzed CpG sites that overlapped with our dataset[108], we were again able to separate these pluripotent cell lines based on our identified hiPSC-specific epigenetic signature. Altogether, these findings indicate that when characterizing the epigenetic differences between hiPSCs and hESCs, cautions must be taken to interpret the results when only a subset of genomic regions is investigated.

Furthermore, we also validated our reprogramming-associated epigenetic signature by using gene expression data from several previously reported datasets[79,

[108]. We observed that a majority of independent clusters separated hiPSC and hESC lines, although clear outliers and different subgroups among hiPSC lines were detected. This result is not totally unexpected because it has been shown that gene expression levels in pluripotent cells are highly variable and depend on how pluripotent cells are generated or maintained[123]. Moreover, Bock et al.[117] also reported the existence of genes in pluripotent cells that contained similar methylation levels but were associated to variable levels of gene expression. Therefore, we cannot exclude the possibility that some hiPSC lines might not segregate well from hESC lines when using the gene expression levels of these nine genes to cluster them.

Finally, although the genes TMEM132D, FAM19A5, and TCERG1L have been reported to be involved in neural processes, we did not identify any significant functional enrichment associated with the nine genes aberrantly methylated in hiPSC lines. Interestingly, Lister et al.[26] identified five of our nine genes (TMEM132C, TMEM132D, FAM19A5, DPP6, and TCERG1L) located within non-CG mega-DMRs as clear outliers in terms of gene expression compared with hESCs. In fact, up to half of their gene outliers located within non-CG mega-DMRs[26] were observed aberrantly methylated in 14 of the 17 hiPSC used in this study. Further studies will be needed to better clarify the role of non-CG mega-DMRs and their implication in the functional behavior of hiPSCs compared with hESCs.

Overall, the results shown here demonstrate the existence of intrinsic common reprogramming-associated epigenetic differences associated with the hiPSC state. We demonstrated that the epigenetic signature described in this work, based on the methylation level of nine genes, can segregate hiPSC and hESC lines in both the

pluripotent state and after differentiation and could explain some of the functional differences between these two pluripotent cell types.

6.6 Acknowledgements

We thank the J.C.I.B. and K.Z. laboratories for helpful discussions and to Dr. Travis Berggren, Margaret Lutz, and Veronica Modesto for their support at the Salk Institute-Stem Cell Core. S.R. was partially supported by Instituto de Salud Carlos III Grant CGCV-1335/07-3. A.G. was supported by the Focht-Powell Fellowship and a California Institute for Regenerative Medicine (CIRM) Predoctoral Fellowship. A.D.P. was partially supported by National Institutes of Health (NIH) Training Grant T32 CA009370. Work in this manuscript was supported by grants from Fundacion Cellex, Ministerio de Economía y Competitividad (MINECO), Sanofi, the G. Harold and Leila Y. Mathers Charitable Foundation, and The Leona M. and Harry B. Helmsley Charitable Trust, California Institute for Regenerative Medicine Grants RB3-05083 and TR1-01273, and NIH Grant R01 GM097253.

Chapter 6, in part, is a reprint of the material as it appears in: Sergio Ruiz*, Dinh Diep*, Athurva Gore, Athanasia D. Panopoulos, Nuria Monsterrat, Nongluk Plongthongkum, Sachin Kumar, Ho-Lim Fung, Alessandra Giorgetti, Josipa Bilic, Erika M. Batchelder, Holm Zaehres, Natalia G. Kan, Hans Robert Scholer, Mark Mercola, Kun Zhang, Juan Carlos Izpisua Belmonte. "Identification of a specific reprogramming-associated epigenetic signature in human induced pluripotent stem cells." *PNAS*. 2012 October 2; 109 (40): 16196-16201. doi:10.1073/pnas.1202352109. Used with permission. The dissertation author was one of the investigators and authors of this paper.

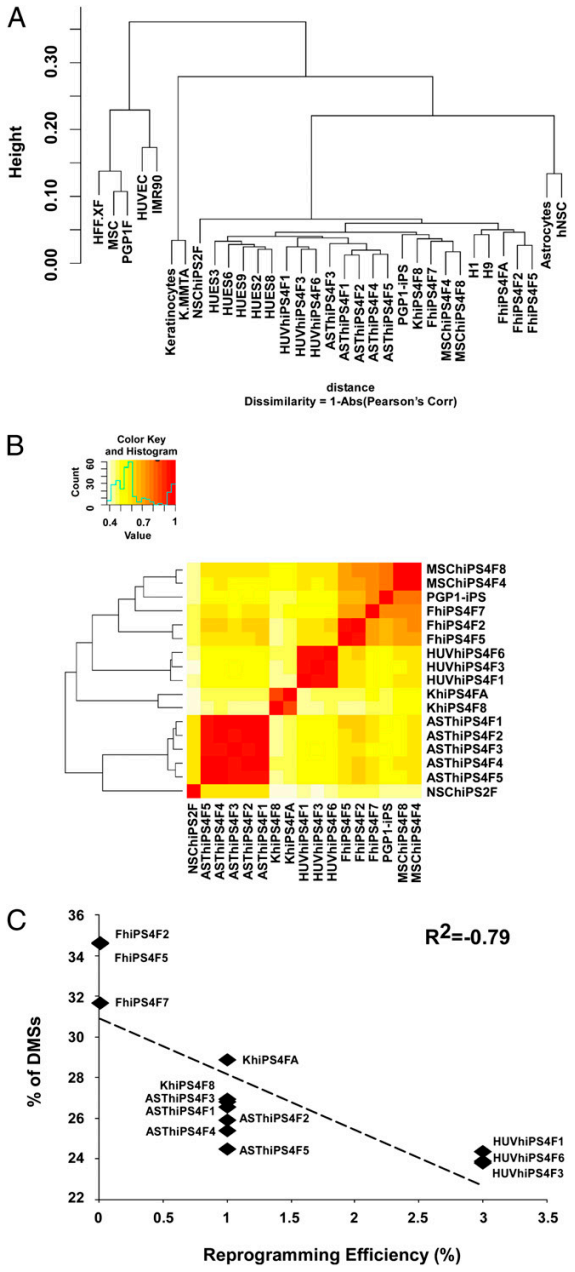


Figure 6.1. Identification and classification of the epigenetic changes occurring during cell reprogramming. (A) Hierarchical clustering of cell lines based on the methylation state of all characterized CpG sites. (B) Heatmap and ordered dendrogram for all hiPSC lines based on the level of relative change observed at each differentially methylated site compared with the values observed in each respective somatic cell of origin. Pearson's correlation values were used to generate a single distance metric. (C) Reprogramming efficiency of somatic cell lines estimated after hiPSCs generation by retroviral infection of OCT4, SOX2, KLF4, and cMYC inversely correlates with the percentage of differential methylation achieved in hiPSC lines. Note that amount of epigenetic reorganization required appears to be a barrier to reprogramming. R^2, Pearson's correlation value.

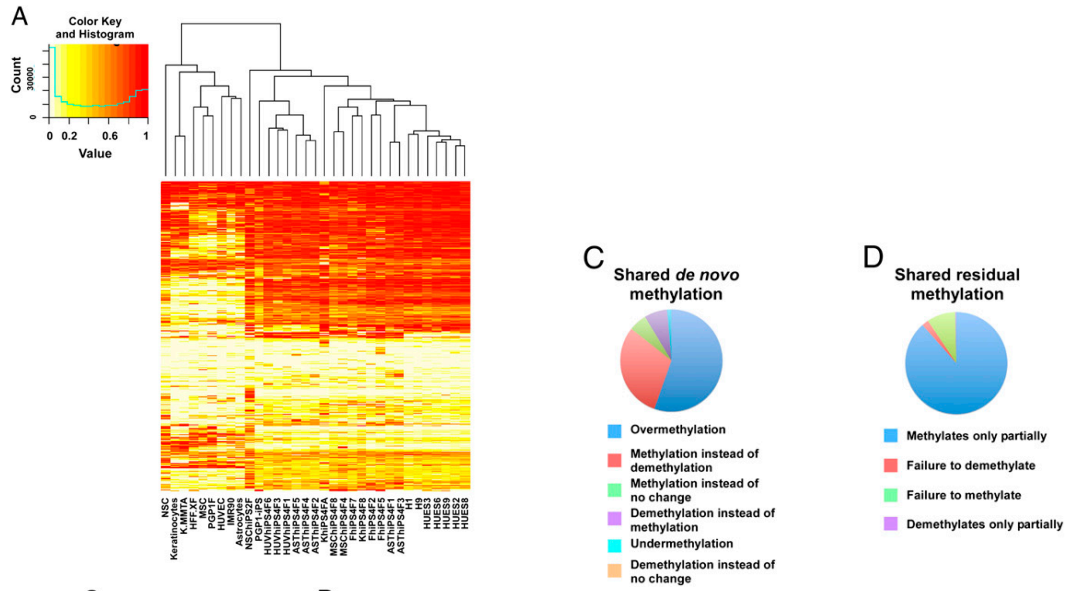


Figure 6.2. Pluripotent cells can be segregated based on the methylation/gene expression level of nine genes. (A) Heatmap and hierarchical clustering results of the cell lines used in this study using methylation patterns at CpG sites containing aberrant methylation in at least one hiPSC line. Similar aberrant epigenetic patterns were observed in hiPSCs derived from common somatic sources, and these lines accordingly tend to cluster together. (B) Graphical representation of an example of residual methylation and de novo methylation located on chromosome 15 (ISLR2 gene). Each circle corresponds to an individual CpG site and the level of methylation is represented in a colored pattern. In the example shown, NSChiPS2F retains the epigenetic pattern of its somatic progenitor (hNSC), showing residual methylation. HUVhiPS4F1 takes on an epigenetic pattern not observed in its somatic progenitor or any of the other pluripotent lines, showing a hiPSC line-specific de novo methylation. Methylation levels of the same CpG sites in hESC and hiPSC lines were included for comparison. (C and D) Types of methylation errors leading to epigenetic aberrations. Most aberrantly methylated CpG sites associated to genes showing de novo methylation (C) and residual methylation (D) in all hiPSC lines are characterized by overmethylation or partial methylation, respectively. (E) Heatmap and ordered dendrogram for the hiPSC and hESC described lines (11) based on the level of relative change observed at CpG sites associated to our nine signature genes. Note that hESC and hiPSC lines segregated in two different groups. (F) Hierarchical clustering of six hiPSC (ASThiPS4F2, 3, 4, and 5, HUVhiPS4F6, and FhiP4F2) and six hESC (H1, H9, HUES2, HUES6, HUES8, and HUES9) lines based on the gene expression level analyzed by real-time PCR of the nine common aberrantly methylated genes identified in hiPSC lines used in this study.

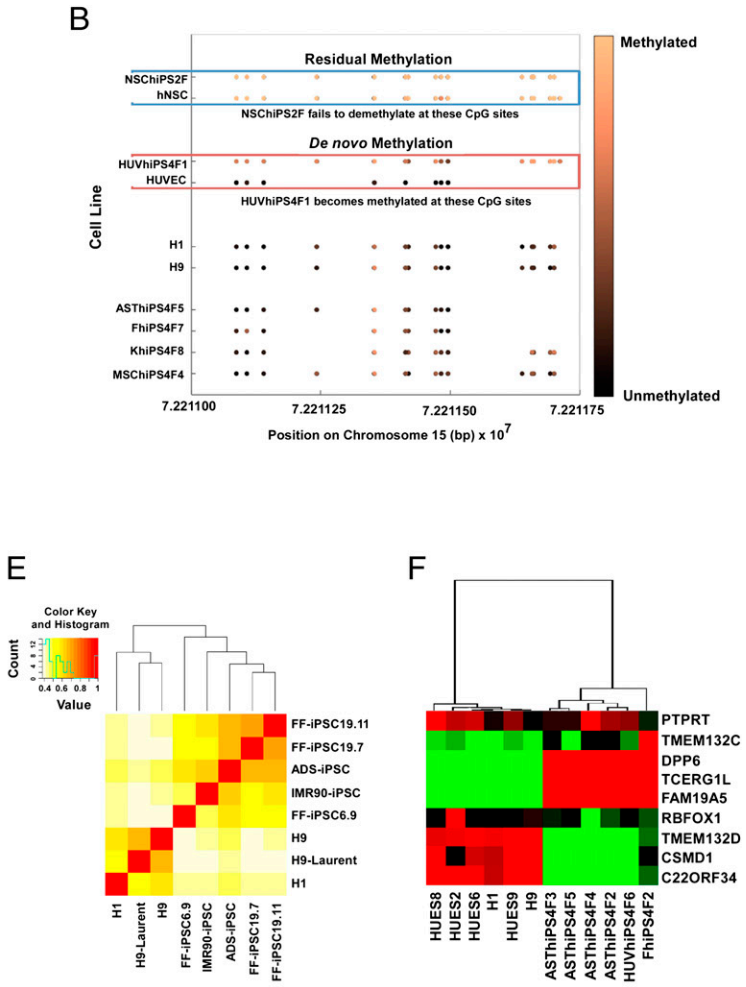


Figure 6.2. Pluripotent cells can be segregated based on the methylation/gene expression level of nine genes (continued).

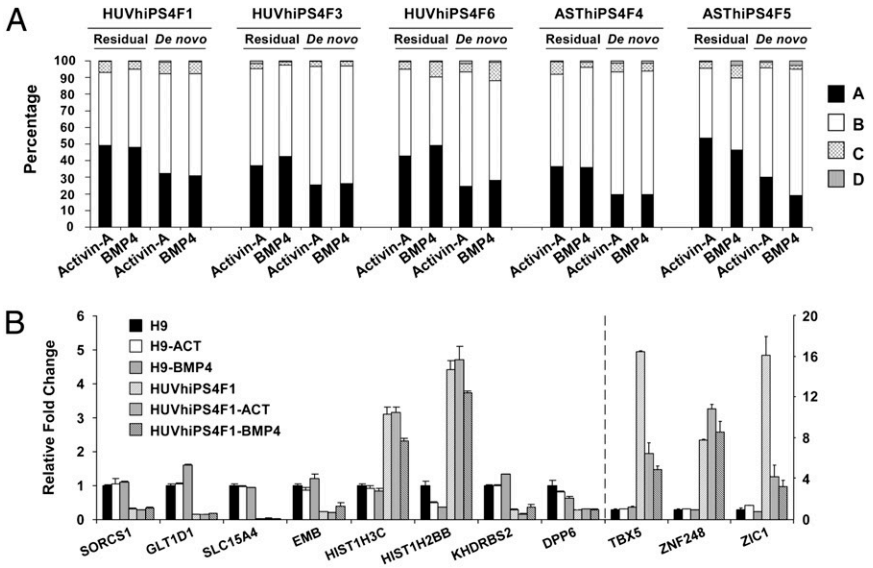


Figure 6.3. Reprogramming-associated epigenetic/transcriptional signatures segregate hiPSCs and hESCs after differentiation. (A) Percentage of aberrant CpG sites identified between hESC-derived lines and the corresponding hiPSC-derived lines classified in the following categories: aberrant methylation remains and is still aberrant compared with differentiated hESCs (A); aberrant methylation remains but is the same as the one found in differentiated hESCs (B); aberrant methylation is removed during differentiation reaching the level found in differentiated hESCs (C); and aberrant methylation changed to a new aberrant methylation state (D). (B) Genes with aberrantly methylated CpG sites and differential transcriptional abundance with at least a twofold cutoff were identified in the HUVhiPS4F1 cell line after comparison with H9 cells. Graph shows the relative fold change in the expression of genes still aberrantly methylated after differentiation between the differentiated HUVhiPS4F1 cell line and the differentiated hESC cell line. Note that differential expression was independent on whether Activin or BMP4-differentiated cells were analyzed. (C) Hierarchical clustering of hESC (H1, H9, HUES3, HUES6, and HUES9) and hiPSC (HUVhiPS4F1, HUVhiPS4F3, HUVhiPS4F6, ASThiPS4F4, and ASThiPS4F5) lines in their pluripotent and differentiated states based on the methylation level of the nine common aberrantly methylated genes identified in the hiPSC lines used in this study. (D) Hierarchical clustering of hESC (H9) and hiPSC (HUVhiPS4F1 and HUVhiPS4F3) lines in their pluripotent and differentiated states based on the gene expression level of the nine common aberrantly methylated genes identified in the hiPSC lines used in this study. Data were obtained from microarray analysis.

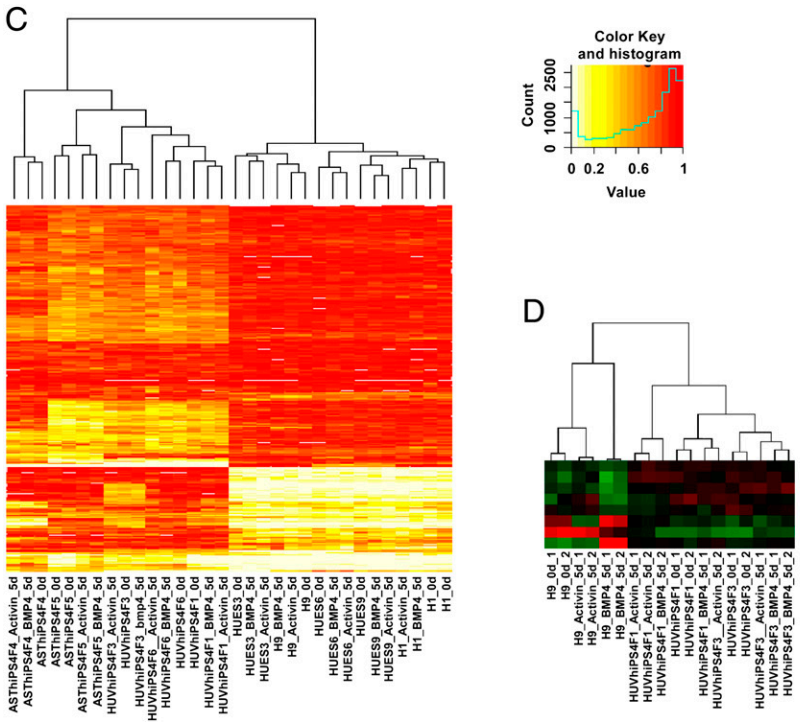
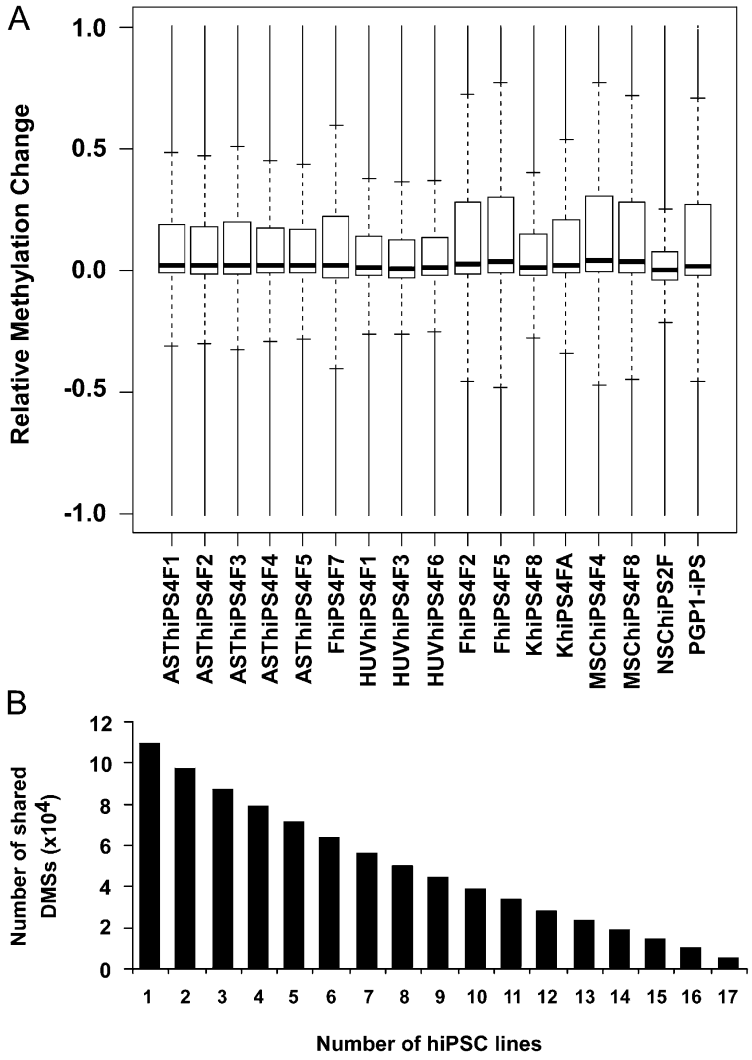


Figure 6.3. Reprogramming-associated epigenetic/transcriptional signatures segregate hiPSCs and hESCs after differentiation (continued).



Supplementary Figure 6.1. Trends observed in differentially methylated sites (DMSs) after reprogramming. (A) Boxplots showing the methylation change observed at all CpG sites in hiPSC lines relative to their somatic progenitors after reprogramming. A −1 value means that a completely methylated site becomes completely unmethylated, whereas a +1 value means that a completely unmethylated site becomes completely methylated. Most CpG sites either do not change in methylation state or become more methylated after reprogramming regardless of somatic progenitor. (B) Cumulative bar plot showing the number of DMSs shared in a defined number of hiPSC lines. For instance, the number of DMSs shared in one hiPSC line represents the number of CpG sites differentially methylated in at least one hiPSC line and the number of DMSs shared in two hiPSC lines represents the number of CpG sites differentially methylated in at least two hiPSC lines. Thus, the number of DMSs shared in all hiPSC lines represents the core number of CpG sites differentially methylated during cell reprogramming regardless of somatic cell source. In these shared DMSs among all of the hiPSC lines, a total of 5,056 CpG sites were hypermethylated, whereas 645 CpG sites were hypomethylated.

Table 6.1. Summary of CpG sites containing residual methylation and de novo methylation in targeted regions

Cell line	Testable sites	% aberrant	% memory	% mutation	No. of genes potentially affected by memory	No. of genes potentially affected by mutation
ASThiPS4F4	434388	1.02	0.45	0.57	191	182
ASThiPS4F5	437266	0.92	0.35	0.58	211	186
ASThiPS4F1	404245	1.30	0.35	0.96	189	310
ASThiPS4F2	380656	1.16	0.41	0.75	171	243
ASThiPS4F3	343025	2.07	0.41	1.65	219	616
FhiPS4F7	340395	2.53	1.27	1.25	487	591
HUVhiPS4F1	374103	1.33	0.38	0.95	200	474
HUVhiPS4F3	392482	1.41	0.42	0.99	251	588
HUVhiPS4F6	433768	1.29	0.32	0.97	190	455
FhiPS4F2	354763	1.62	0.52	1.10	292	213
FhiPS4F5	296451	2.47	0.62	1.85	362	682
KhiPS4F8	396085	2.60	0.82	1.78	586	1040
KhiPS4FA	270126	2.41	0.46	1.95	288	831
MSChiPS4F4	437957	2.34	0.96	1.39	560	462
MSChiPS4F8	429575	2.85	1.60	1.25	896	552
NSChiPS2F	327308	3.82	0.84	2.98	538	1912
PGP1-iPS	437433	2.63	1.47	1.16	997	703

Supplementary Table 6.1. CpG sites targeted in this study are more informative than those targeted in previous studies

Sample	No. of DMSs	DMS, %	Unique DMS
Targeted sequencing of chosen sites			
ASThiPS4F1	45,467	27	82
ASThiPS4F2	44,379	26	103
ASThiPS4F3	46,112	27	262
ASThiPS4F4	43,457	25	58
ASThiPS4F5	41,900	24	110
FhiPS4F2	58,168	35	449
FhiPS4F5	59,266	35	661
FhiPS4F7	54190	32	1,428
HUVhiPS4F1	40,926	24	349
HUVhiPS4F3	40,768	24	366
HUVhiPS4F6	41,697	24	228
KhiPS4FA	49,432	29	1,337
KhiPS4F8	45,924	27	631
MSChiPS4F4	63,976	37	442
MSChiPS4F8	61,842	36	420
NSChiPS2F	38,688	23	2,865
PGP1-iPS	60,820	36	2,286
Targeted sequencing of CpG islands			
BjiPS11	1,354	17	67
BjiPS12	1,437	18	124
IMR90iPS	1,675	21	401
hFib2iPS	1,542	19	386

More differential methylation is observed in the current data set than in previous targeted sequencing experiments. The CpG sites analyzed in this study are therefore more informative than those analyzed in previous studies. Current data is presented in this work. Data obtained from previous studies (1).

Chapter 7: Discussion and Future Directions

The discovery of induced pluripotent stem cells has already begun to revolutionize the field of regenerative medicine. Already, researchers are using iPSCs *in vitro* to culture and study formerly near-inaccessible tissue types, model rare diseases, and test drugs and treatments[21]. Work proceeds on developing transplantable cells and even organs derived from patient-specific iPSCs; in fact, some countries are even beginning to store sets of immunocompatible iPSCs for potential future use in drug development and cell therapies[124]. However, despite the rapid advancement of the field in the last seven years, questions still remain about iPSC safety and efficacy. In order for iPSCs to gain acceptance into the clinic, they must be proven to be both a safe and an effective cell source.

We performed one of the first genomic screens of induced pluripotent stem cells, and identified reprogramming-associated mutations in every cell line examined. The level of mutations was much higher than expected (rivaling that found in cancer cell lines), and remained present regardless of the age of the donor, amount of time progenitor cells were in culture, progenitor cell types used, or reprogramming method used. We identified three distinct categories of reprogramming-associated mutations: pre-existing mutations, which exist at low levels in progenitor cells and are fixed through clonal selection; pre-culture mutations, which occurred early during the reprogramming process in the single cell used to generate the iPSC line; and culture mutations, which occurred during iPSC growth and expansion and became fixed through either selective advantage or random chance. As has been confirmed by other research groups[96, 97], pre-existing mutations comprised a majority of reprogramming-associated mutations. We additionally determined that protein-

coding mutations, when taken individually, do not as a whole seem to have been selected due to their ability to improve reprogramming efficiency.

This might make it seem as if reprogramming-associated mutations are non-functional and therefore not an issue. However, the lack of a common functionality or selection pressure behind reprogramming-associated mutations actually raises a more challenging problem. Because iPSCs will be differentiated into potentially every tissue type, it is possible that mutations could result in unpredictable behavior in alternative tissue contexts. This raises a major potential hurdle for the use of induced pluripotent stem cells; a functional safety test will need to be developed to demonstrate that a given iPSC line will not result in a deleterious phenotype when used in therapy. As several countries are already beginning the process of "banking" iPSC lines with HLA haplotypes that match large portions of the population[125], these banks should implement a functional differentiation-based test in which each chosen iPSC line is differentiated into a desired cell type and thoroughly tested for normal function. Lines could be further classified as safe when differentiated into certain tissue types, and deemed unsatisfactory for others.

Another potential solution to this issue lies in gene correction. Gene correction involves replacement of an unwanted section of DNA, often utilizing the cell's own homologous recombination machinery to replace a mutated region with a clean version[126]. Gene correction has been previously demonstrated to not be an inherently mutagenic process on its own[98]; thus, many iPSC mutations could be corrected. While correcting every mutation in an iPSC line is likely unfeasible, because protein-coding mutations are the most likely to be functional, these reprogramming-associated mutations could be corrected to generate a safer iPSC

line. However, there would still be a possibility that a non-coding mutation that remained in an iPSC line might behave erratically in certain adult cell type contexts.

We also utilized targeted sequencing to characterize the epigenetic state of induced pluripotent stem cells derived from six different cell types. We identified regions of both epigenetic memory (where iPSCs retain the epigenetic state of their progenitor) and aberrant DNA methylation (where iPSCs obtain a unique methylation signature). By utilizing the unique advantages of padlock probes, we were able to target only important genomic regions for analysis, and thereby discover a nine gene iPSC-specific epigenetic signature; this signature was present in the form of tangible gene expression differences in many iPSCs generated in multiple labs. These results demonstrate that most iPSCs contain their own unique gene expression profile that even remains post-differentiation; this must be taken into account when considering iPSC-derived tissue for cell therapies.

We were surprised to learn that this epigenetic signature had not been previously identified in other studies of iPSCs. After investigation, it turned out that two separate factors were behind this issue. First, another sequencing experiment targeting only a portion of the genome through reduced representation bisulfite sequencing (RRBS) did not discover this epigenetic signature due to RRBS only covering CpG-dense regions of the genome[117]. Second, in whole genome bisulfite sequencing experiments, this signature was not discovered due to data normalization; regions containing more CpGs masked the true epigenetic differences[26]. This result further supports targeted sequencing of important genomic regions as an optimal strategy for analyzing epigenetic differences between various cell lines, lest any true differences be masked or missed.

Because of the presence of epigenetic aberrations in induced pluripotent stem cells, a recent study examined if alternative culture conditions could reduce some of these issues. This study[127] identified a set of factors that could cause ESCs and iPSCs to more readily retain full pluripotency in culture; it also determined that somatic cells could be treated with a set of cytokines and small molecule inhibitors during reprogramming to return them to the same more epigenetically "naïve" state. These naïve-iPSCs showed better performance in chimeric experiments and additionally showed a more ESC-like methylation pattern in reduced representation bisulfite sequencing (RRBS) experiments. Because naïve-iPSC gene expression and DNA methylation was globally closer to that of human ESCs, it is possible that the nine gene iPSC-specific epigenetic signature might not be present in these naïve cells. Targeted sequencing of highly informative CpG sites using padlock probes should be performed on these naïve iPSCs to determine if these troubling epigenetic aberrations have truly been eliminated.

Taken together, these results demonstrate that iPSCs must still overcome major hurdles prior to their widespread clinical use. Every single iPSC line contains hundreds of point mutations throughout the genome and misregulation of gene expression that remains even after differentiation. As the field of induced pluripotent stem cells continues to move forward at a record pace, researchers and clinicians must keep in mind that these genomic and epigenomic aberrations are present. Rigorous work towards establishing clinical safety standards for genomic and epigenomic integrity in pluripotent-derived therapies will be essential before the promise of induced pluripotency can be fully realized.

References

1. Mardis, E.R. Next-generation sequencing platforms. *Annu Rev Anal Chem (Palo Alto Calif)* **6**, 287-303 (2013).

2. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931-945 (2004).

3. Dewitt, N.D., Yaffe, M.P. & Trounson, A. Building stem-cell genomics in California and beyond. *Nat Biotechnol* **30**, 20-25 (2012).

4. Linnarsson, S. Recent advances in DNA sequencing methods - general principles of sample preparation. *Exp Cell Res* **316**, 1339-1343 (2010).

5. Bentley, D.R. et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53-59 (2008).

6. Adey, A. et al. Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome Biol* **11**, R119 (2010).

7. Mamanova, L. et al. Target-enrichment strategies for next-generation sequencing. *Nat Methods* **7**, 111-118 (2010).

8. Wang, D.G. et al. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280**, 1077-1082 (1998).

9. Tewhey, R. et al. Microdroplet-based PCR enrichment for large-scale targeted sequencing. *Nat Biotechnol* **27**, 1025-1031 (2009).

10. Porreca, G.J. et al. Multiplex amplification of large sets of human exons. *Nat Methods* **4**, 931-936 (2007).

11. Deng, J. et al. Targeted bisulfite sequencing reveals changes in DNA methylation associated with nuclear reprogramming. *Nat Biotechnol* **27**, 353-360 (2009).

12. Gnirke, A. et al. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotechnol* **27**, 182-189 (2009).

13. Hodges, E. et al. Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel sequencing. *Nat Protoc* **4**, 960-974 (2009).

14. Krueger, F., Kreck, B., Franke, A. & Andrews, S.R. DNA methylome analysis using short bisulfite sequencing data. *Nature Methods* **9**, 145-151 (2012).

15. Robertson, K.D. DNA methylation and human disease. *Nat Rev Genet* **6**, 597-610 (2005).

16. Down, T.A. et al. A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nat Biotechnol* **26**, 779-785 (2008).

17. Li, N. et al. Whole genome DNA methylation analysis based on high throughput sequencing technology. *Methods* **52**, 203-212 (2010).

18. Frommer, M. et al. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc Natl Acad Sci U S A* **89**, 1827-1831 (1992).

19. Pelizzola, M. & Ecker, J.R. The DNA methylome. *FEBS Lett* **585**, 1994-2000 (2011).

20. Thomson, J.A. et al. Embryonic stem cell lines derived from human blastocysts. *Science* **282**, 1145-1147 (1998).

21. Robinton, D.A. & Daley, G.Q. The promise of induced pluripotent stem cells in research and therapy. *Nature* **481**, 295-305 (2012).

22. Noggle, S. et al. Human oocytes reprogram somatic cells to a pluripotent state. *Nature* **478**, 70-75 (2011).

23. Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126**, 663-676 (2006).

24. Zhao, X.Y. et al. iPS cells produce viable mice through tetraploid complementation. *Nature* **461**, 86-U88 (2009).

25. Ghosh, Z. et al. Persistent Donor Cell Gene Expression among Human Induced Pluripotent Stem Cells Contributes to Differences with Human Embryonic Stem Cells. *Plos One* **5** (2010).

26. Lister, R. et al. Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. *Nature* **471**, 68-73 (2011).

27. Yamanaka, S. Patient-Specific Pluripotent Stem Cells Become Even More Accessible. *Cell Stem Cell* **7**, 1-2 (2010).

28. Eckhardt, F. et al. DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat Genet* **38**, 1378-1385 (2006).

29. Cokus, S.J. et al. Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature* **452**, 215-219 (2008).

30. Liu, G.H. et al. Recapitulation of premature ageing with iPSCs from Hutchinson-Gilford progeria syndrome. *Nature* **472**, 221-225 (2011).

31. Liu, G.H. et al. Targeted gene correction of laminopathy-associated LMNA mutations in patient-specific iPSCs. *Cell Stem Cell* **8**, 688-694 (2011).

32. Xu, Y. et al. Genome-wide regulation of 5hmC, 5mC, and gene expression by Tet1 hydroxylase in mouse embryonic stem cells. *Mol Cell* **42**, 451-464 (2011).

33. Wang, H. et al. Rapid identification of heterozygous mutations in Drosophila melanogaster using genomic capture sequencing. *Genome Res* **20**, 981-988 (2010).

34. Becker, R.A., Chambers, J.M. & Wilks, A.R. The new S language : a programming environment for data analysis and graphics. (Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, Calif.; 1988).

35. Hansen, K.D. et al. Increased methylation variation in epigenetic domains across cancer types. *Nat Genet* **43**, 768-775 (2011).

36. Gore, A. et al. Somatic coding mutations in human induced pluripotent stem cells. *Nature* **471**, 63-67 (2011).

37. Turner, E.H., Lee, C., Ng, S.B., Nickerson, D.A. & Shendure, J. Massively parallel exon capture and library-free resequencing across 16 genomes. *Nat Methods* **6**, 315-316 (2009).

38. Shoemaker, R., Deng, J., Wang, W. & Zhang, K. Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome. *Genome Res* **20**, 883-889 (2010).

39. Doi, A. et al. Differential methylation of tissue- and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts. *Nat Genet* **41**, 1350-1353 (2009).

40. Irizarry, R.A. et al. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat Genet* **41**, 178-186 (2009).

41. Lister, R. et al. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**, 315-322 (2009).

42. Figueroa, M.E. et al. DNA methylation signatures identify biologically distinct subtypes in acute myeloid leukemia. *Cancer Cell* **17**, 13-27 (2010).

43. McLean, C.Y. et al. GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol* **28**, 495-501 (2010).

44. Yu, J. et al. Induced pluripotent stem cell lines derived from human somatic cells. *Science* **318**, 1917-1920 (2007).

45. Mayshar, Y. et al. Identification and classification of chromosomal aberrations in human induced pluripotent stem cells. *Cell Stem Cell* **7**, 521-531 (2010).

46. Hong, H. et al. Suppression of induced pluripotent stem cell generation by the p53-p21 pathway. *Nature* **460**, 1132-1135 (2009).

47. Li, H. et al. The Ink4/Arf locus is a barrier for iPS cell reprogramming. *Nature* **460**, 1136-1139 (2009).

48. Kawamura, T. et al. Linking the p53 tumour suppressor pathway to somatic cell reprogramming. *Nature* **460**, 1140-1144 (2009).

49. Utikal, J. et al. Immortalization eliminates a roadblock during cellular reprogramming into iPS cells. *Nature* **460**, 1145-1148 (2009).

50. Marion, R.M. et al. A p53-mediated DNA damage response limits reprogramming to ensure iPS cell genomic integrity. *Nature* **460**, 1149-1153 (2009).

51. Ruiz, S. et al. A high proliferation rate is required for cell reprogramming and maintenance of human embryonic stem cell identity. *Curr Biol* **21**, 45-52 (2011).

52. Bashiardes, S. et al. Direct genomic selection. *Nat Methods* **2**, 63-69 (2005).

53. Akagi, T., Sasai, K. & Hanafusa, H. Refractory nature of normal human diploid fibroblasts with respect to oncogene-mediated transformation. *Proc Natl Acad Sci U S A* **100**, 13567-13572 (2003).

54. Cowan, C.A. et al. Derivation of embryonic stem-cell lines from human blastocysts. *N Engl J Med* **350**, 1353-1356 (2004).

55. Park, I.H. et al. Reprogramming of human somatic cells to pluripotency with defined factors. *Nature* **451**, 141-146 (2008).

56. Chan, E.M. et al. Live cell imaging distinguishes bona fide human iPS cells from partially reprogrammed cells. *Nat Biotechnol* **27**, 1033-1037 (2009).

57. Dimos, J.T. et al. Induced pluripotent stem cells generated from patients with ALS can be differentiated into motor neurons. *Science* **321**, 1218-1221 (2008).

58. Boulting, G.L. et al. A functionally characterized test set of human induced pluripotent stem cells. *Nat Biotechnol* **29**, 279-286 (2011).

59. Rodriguez-Piza, I. et al. Reprogramming of human fibroblasts to induced pluripotent stem cells under xeno-free conditions. *Stem Cells* **28**, 36-44 (2010).

60. Stewart, S.A. et al. Lentivirus-delivered stable gene silencing by RNAi in primary cells. *RNA* **9**, 493-501 (2003).

61. Warren, L. et al. Highly efficient reprogramming to pluripotency and directed differentiation of human cells with synthetic modified mRNA. *Cell Stem Cell* **7**, 618-630 (2010).

62. Aasen, T. et al. Efficient and rapid generation of induced pluripotent stem cells from human keratinocytes. *Nat Biotechnol* **26**, 1276-1284 (2008).

63. Zhang, K. et al. Digital RNA allelotyping reveals tissue-specific and allele-specific gene expression in human. *Nat Methods* **6**, 613-618 (2009).

64. Sakharkar, M.K., Chow, V.T. & Kangueane, P. Distributions of exons and introns in the human genome. *In Silico Biol* **4**, 387-393 (2004).

65. Pleasance, E.D. et al. A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463**, 191-196 (2010).

66. Lee, W. et al. The mutation spectrum revealed by paired genome sequences from a lung cancer patient. *Nature* **465**, 473-477 (2010).

67. Ding, L. et al. Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature* **464**, 999-1005 (2010).

68. Forbes, S.A. et al. The Catalogue of Somatic Mutations in Cancer (COSMIC). *Curr Protoc Hum Genet* **Chapter 10**, Unit 10 11 (2008).

69. Dennis, G., Jr. et al. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol* **4**, P3 (2003).

70. Levy, S. et al. The diploid genome sequence of an individual human. *PLoS Biol* **5**, e254 (2007).

71. Drmanac, R. et al. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327**, 78-81 (2010).

72. Ng, S.B. et al. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**, 272-276 (2009).

73. Kumar, P., Henikoff, S. & Ng, P.C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* **4**, 1073-1081 (2009).

74. Shah, S.P. et al. Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature* **461**, 809-813 (2009).

75. Futreal, P.A. et al. A census of human cancer genes. *Nat Rev Cancer* **4**, 177-183 (2004).

76. Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A. & McKusick, V.A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* **33**, D514-517 (2005).

77. Druley, T.E. et al. Quantification of rare allelic variants from pooled genomic DNA. *Nat Methods* **6**, 263-265 (2009).

78. Ahuja, D., Saenz-Robles, M.T. & Pipas, J.M. SV40 large T antigen targets multiple cellular pathways to elicit cellular transformation. *Oncogene* **24**, 7729-7745 (2005).

79. Yu, J. et al. Human induced pluripotent stem cells free of vector and transgene sequences. *Science* **324**, 797-801 (2009).

80. Takahashi, K. et al. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* **131**, 861-872 (2007).

81. Lowry, W.E. et al. Generation of human induced pluripotent stem cells from dermal fibroblasts. *Proc Natl Acad Sci U S A* **105**, 2883-2888 (2008).

82. Meissner, A., Wernig, M. & Jaenisch, R. Direct reprogramming of genetically unmodified fibroblasts into pluripotent stem cells. *Nat Biotechnol* **25**, 1177-1181 (2007).

83. Hussein, S.M. et al. Copy number variation and selection during reprogramming to pluripotency. *Nature* **471**, 58-62 (2011).

84. Laurent, L.C. et al. Dynamic changes in the copy number of pluripotency and cell proliferation genes in human ESCs and iPSCs during reprogramming and time in culture. *Cell Stem Cell* **8**, 106-118 (2011).

85. Ohi, Y. et al. Incomplete DNA methylation underlies a transcriptional memory of somatic cells in human iPS cells. *Nat Cell Biol* **13**, 541-549 (2011).

86. Kim, J.B. et al. Direct reprogramming of human neural stem cells by OCT4. *Nature* **461**, 649-643 (2009).

87. Panopoulos, A.D. et al. Rapid and highly efficient generation of induced pluripotent stem cells from human umbilical vein endothelial cells. *Plos One* **6**, e19743 (2011).

88. Ruiz, S. et al. High-efficient generation of induced pluripotent stem cells from human astrocytes. *Plos One* **5**, e15526 (2010).

89. Quail, M.A., Swerdlow, H. & Turner, D.J. Improved protocols for the illumina genome analyzer sequencing system. *Curr Protoc Hum Genet* **Chapter 18**, Unit 18 12 (2009).

90. Liu, X., Yu, X., Zack, D.J., Zhu, H. & Qian, J. TiGER: a database for tissue-specific gene expression and regulation. *BMC Bioinformatics* **9**, 271 (2008).

91. Guenther, M.G. et al. Chromatin structure and gene expression programs of human embryonic and induced pluripotent stem cells. *Cell Stem Cell* **7**, 249-257 (2010).

92. Yamanaka, S. Elite and stochastic models for induced pluripotent stem cell generation. *Nature* **460**, 49-52 (2009).

93. Feng, Q. et al. Hemangioblastic derivatives from human induced pluripotent stem cells exhibit limited expansion and early senescence. *Stem Cells* **28**, 704-712 (2010).

94. Hu, B.Y. et al. Neural differentiation of human induced pluripotent stem cells follows developmental principles but with variable potency. *Proc Natl Acad Sci U S A* **107**, 4335-4340 (2010).

95. Miura, K. et al. Variation in the safety of induced pluripotent stem cell lines. *Nat Biotechnol* **27**, 743-745 (2009).

96. Cheng, L. et al. Low incidence of DNA sequence variation in human induced pluripotent stem cells generated by nonintegrating plasmid expression. *Cell Stem Cell* **10**, 337-344 (2012).

97. Young, M.A. et al. Background mutations in parental cells account for most of the genetic heterogeneity of induced pluripotent stem cells. *Cell Stem Cell* **10**, 570-582 (2012).

98. Howden, S.E. et al. Genetic correction and analysis of induced pluripotent stem cells from a patient with gyrate atrophy. *Proc Natl Acad Sci U S A* **108**, 6537-6542 (2011).

99. Woodruff, G. et al. The Presenilin-1 DeltaE9 Mutation Results in Reduced gamma-Secretase Activity, but Not Total Loss of PS1 Function, in Isogenic Human Stem Cells. *Cell Rep* **5**, 974-985 (2013).

100. Shin, H., Liu, T., Manrai, A.K. & Liu, X.S. CEAS: cis-regulatory element annotation system. *Bioinformatics* **25**, 2605-2606 (2009).

101. Pfeifer, G.P., You, Y.H. & Besaratinia, A. Mutations induced by ultraviolet light. *Mutat Res* **571**, 19-31 (2005).

102. Marnett, L.J. Oxyradicals and DNA damage. *Carcinogenesis* **21**, 361-370 (2000).

103. Schuster-Bockler, B. & Lehner, B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature* **488**, 504-507 (2012).

104. Michaelson, J.J. et al. Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. *Cell* **151**, 1431-1442 (2012).

105. Bernstein, B.E. et al. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).

106. Cibulskis, K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* **31**, 213-219 (2013).

107. Daley, G.Q. Common themes of dedifferentiation in somatic cell reprogramming and cancer. *Cold Spring Harb Symp Quant Biol* **73**, 171-174 (2008).

108. Nazor, K.L. et al. Recurrent variations in DNA methylation in human pluripotent stem cells and their differentiated derivatives. *Cell Stem Cell* **10**, 620-634 (2012).

109. Marchetto, M.C. et al. Transcriptional signature and memory retention of human-induced pluripotent stem cells. *Plos One* **4**, e7076 (2009).

110. Bar-Nur, O., Russ, H.A., Efrat, S. & Benvenisty, N. Epigenetic memory and preferential lineage-specific differentiation in induced pluripotent stem cells derived from human pancreatic islet beta cells. *Cell Stem Cell* **9**, 17-23 (2011).

111. Polo, J.M. et al. Cell type of origin influences the molecular and functional properties of mouse induced pluripotent stem cells. *Nat Biotechnol* **28**, 848-855 (2010).

112. Kim, K. et al. Epigenetic memory in induced pluripotent stem cells. *Nature* **467**, 285-290 (2010).

113. Kim, K. et al. Donor cell type can influence the epigenome and differentiation potential of human induced pluripotent stem cells. *Nat Biotechnol* **29**, 1117-1119 (2011).

114. Quattrocelli, M. et al. Intrinsic cell memory reinforces myogenic commitment of pericyte-derived iPSCs. *J Pathol* **223**, 593-603 (2011).

115. Hu, Q., Friedrich, A.M., Johnson, L.V. & Clegg, D.O. Memory in induced pluripotent stem cells: reprogrammed human retinal-pigmented epithelial cells show tendency for spontaneous redifferentiation. *Stem Cells* **28**, 1981-1991 (2010).

116. Chin, M.H. et al. Induced pluripotent stem cells and embryonic stem cells are distinguished by gene expression signatures. *Cell Stem Cell* **5**, 111-123 (2009).

117. Bock, C. et al. Reference Maps of human ES and iPS cell variation enable high-throughput characterization of pluripotent cell lines. *Cell* **144**, 439-452 (2011).

118. Ghosh, Z. et al. Persistent donor cell gene expression among human induced pluripotent stem cells contributes to differences with human embryonic stem cells. *Plos One* **5**, e8975 (2010).

119. Panopoulos, A.D. et al. The metabolome of induced pluripotent stem cells reveals metabolic changes occurring in somatic cell reprogramming. *Cell Res* **22**, 168-177 (2012).

120. Diep, D. et al. Library-free methylation sequencing with bisulfite padlock probes. *Nat Methods* **9**, 270-272 (2012).

121. Xu, R.H. et al. BMP4 initiates human embryonic stem cell differentiation to trophoblast. *Nat Biotechnol* **20**, 1261-1264 (2002).

122. Ruiz, S. et al. Identification of a specific reprogramming-associated epigenetic signature in human induced pluripotent stem cells. *Proc Natl Acad Sci U S A* **109**, 16196-16201 (2012).

123. Newman, A.M. & Cooper, J.B. Lab-specific gene expression signatures in pluripotent stem cells. *Cell Stem Cell* **7**, 258-262 (2010).

124. McKernan, R. & Watt, F.M. What is the point of large-scale collections of human induced pluripotent stem cells? *Nat Biotech* **31**, 875-877 (2013).

125. Turner, M. et al. Toward the development of a global induced pluripotent stem cell library. *Cell Stem Cell* **13**, 382-384 (2013).

126. Naldini, L. Ex vivo gene transfer and correction for cell-based therapies. *Nat Rev Genet* **12**, 301-315 (2011).

127. Gafni, O. et al. Derivation of novel human ground state naive pluripotent stem cells. *Nature* (2013).